DOCUMENT RESUME

ED 386 477 TM 024 031

AUTHOR	Freedle, Roy; Kostin, Irene
TITLE	The Prediction of SAT Reading Comprehension Item Difficulty for Expository Prose Passages. PRPC Final Report P/J 969-60.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-91-29
PUB DATE	Apr 91
NOTE	71p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; *Difficulty Level; High Schools; High School Students; Inferences; Multiple Choice Tests; *Prediction; *Reading Comprehension; Regression (Statistics); Test Construction; *Test Items
IDENTIFIERS	Explicit Instruction; Graduate Record Examinations; *Main Idea; *Scholastic Aptitude Test

ABSTRACT

The primary goal of this project was to examine the predictability of Scholastic Aptitude Test (SAT) reading item difficulty (equated delta) for main idea items, and the predictability of main idea, inference, and explicit statement item types. A secondary purpose was to contrast the responses of high verbal and low verbal ability examinees. Primary attention was paid to studying 110 main idea reading items and their associated passages, but results are also reported for 285 reading items from 34 SAT forms representing a wider range of item types. The percent variance of main idea item difficulty accounted for varied from 46% to 59% depending on the particular analysis. The predictability of all 3 reading item types (n=285) varied from 21% to 29%, depending on the analysis. Results indicated that: (1) multiple-choice reading items are sensitive to variables similar to those reported in the experimental literature on comprehension; (2) many of these variables provide significant independent predictive information in regression analyses; (3) the placement (early versus middle of text) of relevant main idea information affects item difficulty; and (4) considerable agreement between SAT and Graduate Record Examinations reading predictability was found. Nine tables (two in an appendix) present analysis findings. Appendixes contain a glossary, two supplemental tables, and a discussion of scores. (Contains 34 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-91-29

ED 386 477

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE PREDICTION OF SAT READING COMPREHENSION ITEM DIFFICULTY FOR EXPOSITORY PROSE PASSAGES

Roy Freedle
Irene Kostin

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
April 1991

TM024031

The Prediction of SAT Reading Comprehension Item Difficulty for Expository Prose Passages

Roy Freedle and Irene Kostin

PRPC Final Report P/J 969-60

Copyright © 1991. Educational Testing Service. All rights reserved.

Abstract

The primary goal of this project was to examine the predictability of SAT reading item difficulty (equated delta) for main idea items, and collectively, the predictability of three major reading item types: main idea, inference and explicit statement items. A secondary purpose in predicting item difficulty was to contrast the responses of high verbal and low verbal ability examinees. Primary attention was paid to studying 110 main idea reading items and their associated passages. However, additional results are reported for 285 reading items taken from 34 disclosed SAT forms which represented a wider range of reading item types.

The percent variance of main idea item difficulty accounted for varied from 46% to 59% depending upon the particular analysis. The predictability of all three reading item types (n = 285) varied from 21% to 29%, depending upon the particular analysis.

Details of item predictability were explored by evaluating several hypotheses. Results indicated that (1) multiple-choice reading items are sensitive to variables similar to those reported in the experimental literature on comprehension, (2) many of these variables provide significant independent predictive information in regression analyses, (3) the placement (early versus middle of text) of relevant main idea information affects item difficulty, and (4) considerable agreement between SAT and GRE reading predictability was found. Additional results contrast the performance of high and low ability groups.

Introduction

Purpose of current study

The purpose of the current study is to predict reading item equated delta values for each of three SAT reading item types: main ideas, inferences and explicit statement items which together constitute about 75% of the reading items. The primary focus is on main idea items. A secondary concern is to compare the predictability of high and low performing examinees.

Background studies

Only a few studies appear to have focused on predicting item difficulty using items from standardized ability tests (Drum, Calfee, & Cook, 1981; Embretson & Wetzel, 1987). While not specifically focused on predicting reading item difficulty, many other studies of language processing have isolated a wide variety of variables that are known to influence comprehension difficulty with respect to decision time and recall measures. A few such studies of particular interest here are the study of negations by Carpenter and Just (1975), the study of rhetorical structure (Grimes, 1975) and its effect on accuracy of prose recall (Meyer, 1975; Meyer & Freedle, 1984) and prose comprehension (Hare, Rabinowitz, & Schieble, 1989), the use of referential expressions in constructing meaning (Clark & Haviland, 1977), and the use of syntactic "frontings" (see details below) which appear to guide the interpretations of semantic relationships within and across paragraphs (see Freedle, Fine, & Fellbaum, 1981). The particular manner in which these selected variables will be studied will become evident later in this report. Using this set of presumably relevant variables, the primary aim of this study has been to try to capture the large- and small-scale structures of prose, and their associated items, in order to best account for observed reading item difficulty in a multiple-choice testing context.

First we review those studies that predict reading item difficulty for multiple-choice tests.

Drum, Calfee, and Cook (1981) predicted item difficulty using various surface structure variables and word frequency measures for the text, and several item variables which also depended on surface structure characteristics (e.g., number of words in the stem and options, number of words with more than one syllable, etc.). They reported good predictability using these simple surface variables; on average, they indicate that about 70% of the variance of multiple-choice reading item difficulty was explained. However, while the Drum et al. (1981) study was innovative in analyzing the multiple-

choice testing process into its constituent parts (i.e., determining the relative contribution of the items, stems, the item's correct and incorrect options as well as the text variables to item difficulty), some of the study's analyses appear to be flawed. Ten predictor variables were extracted from very small reading item samples (varying between 20 and 36 items) taken from seven children's reading tests. At most one or two predictors instead of 10 should have been extracted from such small samples--see Cohen and Cohen (1983); hence, 70% of the item difficulty variance is probably too large an estimate of the variance actually accounted for.

Embretson and Wetzel (1987) also studied the predictability of 75 reading item difficulties using a few of the surface variables studied by Drum et al. (1981). But in addition, because of the brevity of their passages, Embretson and Wetzel (1987) were able to do a propositional analysis (see Kintsch & van Dijk, 1978) and add variables from this analysis, along with several other measures, as predictor variables. In particular they found that connective propositions were significant predictors. We believe that Meyer's (1975) top-level rhetorical structures, which we include in the present study, indirectly assess the presence of connectives (such as and, but, however, since, because, etc.) since each of the rhetorical devices differently emphasizes these connectives. For example, a top-level causal structure tends to use connectives such as since and because. A list structure tends to use connectives such as and and then, while a comparative structure will often employ connectives such as however, yet, etc.

Now we quickly review those additional studies which deal with variables that have been found to influence reading comprehension difficulty. Most of these additional variables were investigated in empirical studies which did not use multiple-choice methods to yield an index of comprehension difficulty. Instead many used dependent measures such as recall of passages or decision time to infer the influence that certain variables have on comprehension difficulty. This review along with our earlier review of the Drum et al. (1981) and Embretson & Wetzel (1987) studies will help us to select a final set of variables which we postulate may also index comprehension difficulty within a multiple-choice testing format.

Carpenter and Just (1975) found that the occurrence of sentence negation increases comprehension decision time. This suggests that the number of negations contained in SAT reading passages may also influence multiple-choice item difficulty. Furthermore, one can inquire whether additional negations that are used in the item structure itself (either in the item stem and/or among the response options) may also separately contribute to comprehension difficulty over and above the contribution of text negations.

Abrahamsen and Shelton (1989) demonstrated improved comprehension of texts that were modified, in part, so that full noun phrases were substituted in place of referential expressions. This suggests that texts with many referential expressions may be more difficult than ones with few referential expressions. Again, for purposes of studying more broadly the effect of number of referential expressions on comprehension difficulty of multiple-choice tests, a separate count is also made of referential expressions that occur in the item proper.

Hare et al. (1989) studied, in part, the effect of four Grimes' (1975) rhetorical organizers on difficulty of identifying the main idea of passages--students either wrote out the main idea if it wasn't explicitly stated or underlined it if it was explicitly stated. They found a significant effect of rhetorical organization such that list type structures (see definitions and examples below) facilitated main idea identification whereas some non-list organizers made main idea information more difficult to locate. Meyer and Freedle (1984) examined the effect of the Grimes (1975) organizers on the ability of students to recall passages which contained the same semantic information except for their top rhetorical organization. They found, like Hare et al. (1989), that list structures facilitated recall (for older subjects). However, they also reported that university students were best helped by comparative type organizations; this latter finding was not replicated by Hare et al. (1989).

It seems likely that rhetorical organization will contribute to comprehension difficulty within a multiple-choice testing format; however, it is not clear, given the differences between Meyer and Freedle (1984) and the Hare et al. (1989) studies, whether we can say in advance which type structure will be found to facilitate performance. Top level rhetorical structure meaningfully applies only to the text structure; a comparable entry for items is not feasible.

Freedle, Fine, and Fellbaum (1981) report differences in the use of "fronted" structures at sentence beginnings (and paragraph beginnings) as a function of the judged quality of student essays. Fronted structures included the following: (1) Cleft structures ("It is true that she found the dog", where the initial "it" is a dummy variable having no referent), (2) marked topics consisting of several subtypes (a) opening prepositional phrases or adverbials ("In the dark, all is uncertain"; "Quickly, near the lodge, the boat overturned") or (b) initial subordinate clauses ("Whenever the car stalled, John would sweat") and (3) combinations of coordinators and marked topics or cleft structures that begin independent clauses ("But, briefly, this didn't stop him"; "And, furthermore, it seems that is all one should say").

Freedle et al. (1981) showed that these different fronting structures significantly discriminate among essay quality such that the better essays contained a higher mean frequency of each of these fronted structures even after partialling out the effect of different lengths of essay as a function of ability level. They interpreted these fronted structures as authors' explicit markers for guiding readers to uncover the relationships that exist among independent clauses. It is not immediately clear whether differential use of all such structures would itself facilitate or inhibit comprehension of SAT passages. If we assume that the structures produced by the more able writers are structures that are more difficult to learn, then one can predict that the more frequently these fronted structures occur, the more difficult the text should be to understand. In support of this, Clark and Haviland (1977) suggest that at least cleft structures may be harder to understand than simple declarative sentences. Also Bever and Townsend (1979) found that when main clauses follow a subordinate clause such sentences are more difficult to process than when main clauses occur in initial sentence positions (this overlaps somewhat with frontings, since initial subordinate clauses would count as one type of fronting). By including a count of all such variables we can explicitly test the relevance of clefts and other fronted structures for how they might affect comprehension difficulty in a multiple-choice testing context. This is done separately for text as well as item content.

Other variables that we hypothesize will be of importance in affecting comprehension difficulty for multiple-choice tests are: vocabulary level (Graves, 1986), various measures of sentence complexity such as sentence length (Klare, 1974-1975), paragraph length (Hites, 1950), number of paragraphs (Freedle, Fine, & Fellbaum, 1981) and abstractness of text (Paivio, 1986). In particular, less frequently occurring words and longer sentence structures tend to make texts more difficult to understand, as can be inferred from their use in traditional readability formulas (Graves,1986); in addition, longer paragraphs, and abstractness of texts also make passages more difficult to comprehend [see Hites (1950) and Paivio (1986), respectively]. Use of more paragraphs was found to be positively correlated with the quality of written essays (Freedle, Fine, & Fellbaum, 1981); it remains to be seen whether number of paragraphs itself contributes to reading comprehension difficulty in a multiple-choice testing context.

Hence one of the hypotheses which we seek to confirm in the present study is that many of the above-mentioned variables which are known to contribute to comprehension difficulty in non-multiple-choice testing formats (or to quality judgments of written essays) will be found to significantly

affect comprehension measures as determined within a multiple-choice testing format. Stating this more succinctly we have:

Hypothesis 1. The following variables significantly influence reading item difficulty as determined within a multiple-choice testing format:

a. negations
b. referentials
c. rhetorical organizers
d. fronted structures:
   1. cleft-structures
   2. marked-topics
   3. combinations (of coordinators and marked topics or coordinators with cleft structures)
e. vocabulary
f. sentence length
g. paragraph length
h. number of paragraphs
i. abstractness of text

Hypothesis 1 is not necessarily a trivial hypothesis at least insofar as the above variables are seen to apply to the coding of the reading passage. Royer (1990) claims that "There is evidence that standardized reading comprehension tests that utilize multiple-choice questions do not measure the comprehension of a given passage. Instead they seem to measure a reader's world knowledge and his or her ability to reason and think about the contents of a passage" (Royer, 1990, p. 162). Royer (1990) then cites work by Tuinman (1973-1974), Drum et al. (1981) and Johnston (1984) to bolster this claim. Tuinman's work is similar to the findings of Katz et al. (1990) wherein multiple-choice reading items are correctly responded to above chance levels in the absence of the reading passage. Of course Katz et al. (1990) also show that a significant increase in correct responses occurs when the passage is subsequently made available to the students. Hence it seems that Royer (1990) appears to have overgeneralized the importance of just item structure in concluding that multiple-choice reading tests do not measure passage comprehension. That is, if multiple-choice tests of reading did not tap passage comprehension and were solely a reflection of world knowledge and reasoning ability, then the subsequent addition of the passage should have had no noticeable effect on reading item correctness. Since Katz et al. (1990) clearly showed a significant augmentation of item correctness when the passage was available one must conclude that multiple-choice reading tests do measure passage comprehension and simultaneously tap other abilities such as reasoning.

Royer's (1990) citation of Drum et al. (1981) also concerns the claimed importance of just item structure to reading comprehension item correctness. Incorrect option plausiblity was the most important predictor in Drum et al.'s (1981) study. They classified this as an item variable. However we claim that incorrect option plausibility is more accurately classified as a text by item interaction, and is not just an item variable. That is, in order to decide whether an incorrect option is a plausible answer or not, one necessarily must scan not only the item information but the text information as well. Hence Drum et al.'s best predictor is one that necessarily implicates the reading of the text. This leads us to conclude that Royer's (1990) acceptance of Drum et al.'s (1981) classification scheme led him to use their results, incorrectly we feel, to support further his hypothesis that text comprehension does not play a crucial role in multiple-choice reading tests.

But suppose Royer's critique of multiple-choice tests is assumed to be correct. Then there is little reason to expect that the nine variables listed under hypothesis 1 (a through i above at least as it applies to the coding of the text) will be significantly related to multiple-choice reading test item difficulty. This should follow because, by hypothesis, multiple-choice tests are not tests of passage comprehension; hence variables (as assessed for the passage) which are known to be related to comprehension difficulty (in the experimental literature), should not correlate with performance on multiple-choice reading comprehension tests. However, if Royer (1990) is incorrect, then there is good reason to suppose that most if not all of the nine variables listed under Hypothesis 1, at least as applied to the coding of the text, will be found to significantly correlate with reading item difficulty as obtained from multiple-choice testing.

If supporting evidence is found for hypothesis 1, there is a second implication that is important to evaluate. There are few studies that simultaneously assess the influence of many variables on comprehension (Goodman, 1982). Furthermore, many of the text materials which are evaluated in the experimental literature are not naturalistic texts but rather are artificially constructed to test the effect of one or two variables (see Hare et al., 1989). With the current SAT passages which are selected from naturalistic texts, it should be possible to evaluate via regression analyses whether the nine categories of variables of Hypothesis 1 contribute independent information in accounting for reading comprehension item difficulty. This leads us to our second hypothesis:

Hypothesis 2. Many of the nine categories of variables provide independent predictive information in accounting for reading item difficulty.

Corollary to Hypothesis 2. Confirmation of hypothesis 2, using SAT data, implies that many of the nine categories of variables for hypothesis 1 apply to naturalistic texts as well as to the more controlled texts employed in many experimental studies of reading comprehension.

There is one last implication that can be tested if Royer (1990) and some portions of the Katz et al. (1990) results are correct--the portions which led them to conclude that multiple-choice reading tests are not valid measures of passage comprehension because items can be responded to above chance levels of correctness in the absence of reading the passage. One can infer that item variables alone must be more important predictors of item difficulty than are text and text associated variables. This leads us to our third hypothesis.

Hypothesis 3a. Item variables alone account for item difficulty variance; text variables do not provide additional predictive information. [Based on implications of Royer (1990) and the conclusions reached by Katz et al. (1990).]

However, if, as we suspect, the evidence shows that multiple choice tests of reading comprehension do measure passage comprehension, then text variables should be found to be significant predictors of item difficulty even after the effect of just the item predictors has already been extracted. Hence we would not be surprised if Hypothesis 3a is not supported.

Other variations on Hypothesis 3a are easy to state.

Hypothesis 3b. Item variables alone account for item difficulty variance; text plus text by item interaction variables do not provide additional predictive information.

Hypothesis 3b will almost certainly not be supported since the Drum et al. (1981) study shows that at least one text by item variable is a good predictor of item difficulty. We state it separately here primarily to clarify statements in the literature (e.g., Royer, 1990) which we feel have incorrectly conflated item and at least one text by item interaction variable into a single category, that of item variables. Hypothesis 3b predicts that none of the text and none of the text by item interaction variables will provide independent predictive information after the effect of item variables have been partialled out.

Hypothesis 3c. When the proportion of variance accounted for by item versus text variables is compared, the contribution of item variables will

always be larger than the contribution of text variables.

Hypothesis 3d. When the proportion of variance accounted for by item versus text plus text by item interaction variables is compared, the contribution of item variables will always be larger than the contribution of text and text associated variables.

Hypotheses 3c and 3d simply make more explicit that even if some variance is contributed by the text (and/or the text associated) variables, that the proportion accounted for by the item variables will always be larger.

Again, it should be clear that we expect all these variants of Hypothesis 3 to receive no support using the SAT reading data. They are stated as they are in order to honor the conclusions reached in some of the published literature on multiple-choice tests of reading comprehension, especially the work of Royer (1990) and Katz et al. (1990).

Background studies for the study of main idea variables. Kieras (1985) specifically focused on the perception of main idea information in reading. Kieras (1985) examined, in part, how students perceived the relative location of main idea information in short paragraphs. He found, using single paragraph passages extracted from technical manuals, that most students perceive main idea information as located early in the paragraph, a few thought the main idea occurred at or near the end of the paragraph, while information in the middle of the paragraph was the least often perceived as a statement of the main idea. Kieras (1985) did not report the relative frequencies with which the actual main ideas occurred among the passages so it is difficult to determine whether students tend to select the opening sentences of passages as containing the main idea because most of the passages placed the key idea in this place or whether the students were simply reflecting a response bias to choose the opening sentences. Unless the main idea was equally represented by its location across the stimulus passages, the Kieras results are ambiguous.

However, the work of Hare et al. (1989) helps to clarify this issue. In one of their studies they systematically varied the known location of a main idea sentence in three locations: the opening sentence, the medial sentence or the final sentence of a paragraph. The experimental subjects underlined which sentence they thought was the main idea sentence. Correct identifications were greatest for initial occurrence of main idea sentences. One can infer from the Hare et al. (1989) results that two tendencies contribute to the main idea correctness: opening sentences that do contain the main idea tend to be selected partly because of a prior bias to select early sentences, but also

because students are attempting to understand the information in the text sentences.

One can generalize the Hare et al. (1989) work including the Kieras' (1985) findings to generate hypotheses concerning how students will respond to multiple-choice items regarding the location of main idea information as it applies to multiparagraph passages. In addition it is not clear whether Kieras' (1985) findings can be generalized to nontechnical as well as technical prose.

If students tend to perceive early text information, especially information in the opening sentences of the first paragraph, as main idea information, then when certain passages actually confirm this search strategy, such items should be easier than those that disconfirm it (where disconfirming main idea information would be information that occurs in the middle of a multi-paragraph text or that occurs primarily at the end; it is disconfirming only because it fails to conform to the expectation that main idea information "should" be near the beginning of a passage). So, the relative ordering of difficulty should be: opening sentences that fit the main idea information as stated in the correct answer to a main idea item will be easiest (other things being equal), while main idea information that occurs near the middle of a text will be associated with the hardest main idea items.

Summarizing the comments above we have the following additional hypothesis to be evaluated.

Hypothesis 4. Relevant main idea information that is located early in the passage will facilitate main idea item correctness; relevant main idea information that is located in the middle of the passage will lead to poorer performance in correctly responding to main idea items.

If supporting evidence is found for this hypothesis this implies that Kieras' (1985) result generalizes to multiple-choice and multi-paragraph contexts of evaluation for both technical as well as nontechnical passages.

Abelson and Black (1986) have contrasted three models for representing text information: the propositional approach (Kintsch, 1974; Kintsch & Van Dijk, 1978); the text grammar and top-rhetorical analysis approach (Grimes, 1975; Meyer, 1975; Mandler & Johnson, 1977); and the content-functional approach which emphasizes why a passage has been written (the functions that it serves; the 'point' that the author is trying to make). Abelson and Black (1986) illustrate how the same passage (usually a short prose selection) can be represented for each model. More importantly, they also illustrate how the

exact phrasing of different questions about the passage can be made to favor one model over another. Hence the implication of their work suggests that multiple-choice formats as they are currently constructed (i.e., constructed without reference to any particular text processing theory) cannot be used to evaluate which text representation process is optimal.

Materials and Method

Each SAT form contains six reading passages. Associated with each passage is a variable number of items, usually between three and five items. A total of 25 reading items is associated with these six passages.

The statistics for each reading item are tabulated from a random sample of approximately 1500 examinees. Furthermore, the sample of approximately 1500 examinees is divided into five ability levels depending upon their total verbal SAT score. Separate statistics (percent pass) are provided for each of the five ability levels for each item. The item statistics card also includes the equated delta for each item. Because only a percent pass is available for each ability level, a z-score transformation had to be determined for each item for the highest and lowest ability levels (only these two extreme ability levels were analyzed).

A sample of 285 reading comprehension items, taken from 34 disclosed SAT forms, comprise the total item sample. All the available disclosed forms that had easily accessible item statistics were considered for inclusion. The total number of reading passages represented was 110. Only main idea (n=110), inference (n=97) and explicit statement (n=78) items were selected for study. One main idea item was used per reading passage. If a passage did not contain a main idea item it was not included in this sample. All inference and explicit statement items (except for those special types listed below) associated with these main idea item passages were also included in the sample.

Examples of each of the three reading item types are:

Main Idea: "The central purpose of the passage is to:

(a) announce the discovery of a great artist
(b) describe and analyze a work of art
(c) point out the historical inaccuracies of a painting
(d) provide an example of the pastoral school of landscape painting
(e) criticize the behavior of the Spanish in the New World"

Inference: "It can be inferred from the passage that Milton believed that Parliaments moral responsibility to the English public was to:

(a) lead by its good example
(b) control major corrupting influences
(c) dictate public morality through noncoercive means

(d) punish only individuals who defy the law
(e) allow the public full freedom in moral matters"

Explicit statement: "According to the passage Black representation in the New York State Assembly before 1920 was hampered by the:

(a) solidly residential nature of the Black community
(b) indifference of other ethnic groups
(c) division of the Black vote between two districts
(d) inability of Black voters to agree on candidates
(e) failure of Harlem voters to sponsor candidates."

Other item types that inquire about an author's tone (e.g., use of irony) and author's organization (e.g., in asking how the first paragraph is related to the second) occur less often and were not scored. We also did not sample items which use a Roman Numeral type format [e.g., where different combinations of 3 elements comprise the list of options as in (a) only I is correct, (b) only I and II are correct, (c) I and III are correct, (d) II and III are correct (e) none are correct]. We also excluded special items which featured a capitalized NOT or LEAST in the item stem. Narrative passages were excluded from this analysis because we focused on just expository type prose. [Narrative passages can be excerpts from novels or short stories (e.g., a passage from Pride and Prejudice).]

Independent variables for representing item, text, and text by item information

The variables are grouped below according to whether they are associated with the items, the text, the text associated variables or whether they are dependent variables. The glossary in Appendix A provides another listing of these variables in ascending numerical order.

Item variables

Item Type

v60 Item type: Main idea
v61 Item type: Inference
v62 Item type: Explicit statement

Variables for item's stem

v14 Stem: Number words in stem (the item question)
v68 Stem: Use of hedge (e.g., perhaps, probably) in stem
v69 Stem: Use of full question or sentence fragment

v70 Stem: Use of simple negation

v71 Stem: Use of fronting (e.g., use of any phrases or clauses preceding the subject of the main independent clause, or use of clefts--see below under text variables for details)

v72 Stem: Sum of referentials to text, stem or options (see below for definitions under text variables)

v73 Stem: Reference made to text lines or paragraphs

Variables for item's correct option

v3 Correct: Ordinal position of correct answer

v15 Correct: Number words in correct option

v75 Correct: Frequency of simple negations in correct option

v76 Correct: Use of fronting in correct option (presence or absence)

v77 Correct: Frequency of referentials in correct option

Variables for item's incorrect options

v16 Incorrects: Number words in all incorrect options

v78 Incorrects: Frequency of simple negations in all incorrect options

v79 Incorrects: Frequency of frontings in all incorrect options

v80 Incorrects: Frequency of referentials in incorrect options

Text Variables

Vocabulary variable for text

v17 Number of words with three or more syllables for the first 100 words of the passage (estimates vocabulary difficulty)

Concreteness/abstractness of text

v44 Is main idea of text and its development basically concerned with concrete or abstract entities?

Subject matter variables of text[1]

v31-v35 The type of semantic content of passage

v31, physics

v32, biology

(v31 and v32 were combined to represent the natural science category--v100)

v33, social sciences

v34, humanities

v58, represents an excerpt of natural science

v59, represents a passage about natural science

Type of rhetorical organization

v35, argumentative passage (i.e., author favors one of several points of view presented in text; occasionally other viewpoints only may be implied)

v45-v48 Grimesean type of rhetorical organization of passage.

v45: List (and/or describe) interrelates a collection of elements in a text which are related in some unspecified manner; a basis of a list "... ranges from a group of attributes of the same character, event, or idea, to a group related by simultaneity to a group related by time sequence" (Meyer, 1985, p. 270). Describe relates a topic to more information about it. We felt this was sufficiently similar to list to warrant scoring them as members of the same category.

---

[1]The six content areas listed in the SAT are: physics, biology, social science, humanities, argue, and narrative. The first four categories contain only expository prose. These four categories are mutually exclusive. The "argue" category, by contrast, reflects a rhetorical structure: the author of the passage is biased towards one viewpoint of the several presented--this represents a positive instance of the "argue" category; the absence of "argue" would be where an author represents one viewpoint, or more viewpoints, with an equal weight given to each. Clearly, one can have a biology passage which is either argumentative or not; this is true for other expository materials as well. Narrative structure represents a different discourse genre and so has not been included in our sample. Note that "argue" partially overlaps with v337 which is a comparative-adversative rhetorical organizer.

v46: Causals "... shows a causal relationship between ideas where one idea is the antecedent or cause and the other is a consequent or effect. The relation is often referred to as the condition, result or purpose with one argument serving as the antecedent and the other as the consequent. The arguments are before and after in time and causally related." (Meyer, 1985, p. 271).

v47: Compare. The comparison relation points out differences and similarities between two or more topics. The two subtypes used here are v337 (compare-adversative which relates a favored view to a less desirable opposing view), and v339 (comparison-alternative which interrelates equally weighted alternative options or equally weighted opposing views). (Meyer, 1985, p. 273).

v48: Problem/solution is defined as follows: "similar to causation in that the problem is before in time and an antecedent for the solution. However, in addition there must be some overlap in topic content between the problem and solution; that is, at least part of the solution must match one cause of the problem. The argument (e.g., problem and solution) are equally weighted and occur at the same level in the content structure." (Meyer, 1985, p. 272).

Coherence of lexical concepts over whole text

v4 Coherence: this involves judging whether opening concepts of the first sentence occur throughout text 3= maximum lexical coherence, . . . 0= no obvious lexical overlap

Lengths of various text segments

v11 = Number paragraphs
v12 = Number words
v13 = Number sentences
v18 = Number words in first paragraph
v19 = Number words in longest paragraph
v42 = Number of sentences in first paragraph
v43 = Number of sentences in longest paragraph

v89 Average number of words per sentence
v90 Average number of words per paragraph
v96 Average length of sentences in first paragraph
v97 Average length of sentences in longest paragraph

Occurrence of different text "frontings" (v50-v55,v57 Use of "frontings" of different types. Some examples follow.

Use of theme-marking: In the front, the car rocked.
Fortunately, the car rocked.
Use of coordination: But, the car rocked.
Use of deferred foci: It is too bad. There are many. (These are "clefts" that function as dummy sentence variables.)
Use of combinations: And, near the rear, the toy fell.
Longest run of frontings: Number of successive independent clauses which begin with fronted information: e.g., "The man laughed. Then, he frowned. And when he turned, fell." This example of three independent clauses has two successive sentences with fronted material; hence its run length is '2'.

v50 - percent fronted clauses, paragraph opening clauses
v51 - frequency fronted clauses, paragraph opening clauses
v52 - percent fronted clauses, total text
v53 - frequency fronted clauses, total text
v54 - frequency combinations of fronted structures, total text
v55 - frequency deferred foci (one type of fronting)
v57 - number of longest run of consecutive fronted clauses

Number of text questions

v56 Number of text questions

Text referentials

v63 - frequency within clause referentials
e.g., "When George fell, he hurt."
v64 - frequency across clause referentials
e.g., "George fell. That hurt."
v65 - frequency special referentials (reference outside text); e.g., "One might feel sorry for George."
v66 - sum of v63,v64,v65

Text negations

v67 Number simple negations in text

Special Text by Item Interaction Variables

The location of text information relevant to answering main idea items correctly.

Variables v6,v37,v39-v41,v86,v87,v338 specify location of main idea in surface text (the following are all dichotomous variables):

v86 - main idea in first sentence of text
v87 - main idea in second sentence of text
v39 - main idea in first short paragraph (100 words or less)
v40 - main idea in first sentence of 2nd paragraph
v338 - main idea is near middle of passage
v6 - main idea in last short paragraph (100 words or less in paragraph)
v41 - main idea is in last sentence of text
v37 - main idea is not located in any specific part of the text

[several of the analyses below used a combined category i.e., v342=v86+v87+v39 since this improved predictability of some of the criterion variables]

Dependent variables: general orientation

For several analyses, the dependent variable of interest was an item's equated delta (an item's difficulty which converts percent corrects per test form to a common scale with mean 13.0 and S.D. of 4). Each item's delta is based on the responses of approximately 1,500 students who are randomly selected from the population that takes a particular SAT form.[2] The equated delta allows one to combine data across test forms by smoothing out small differences in item difficulty that occur because some test forms are taken by slightly higher ability examinees at different times of the year. It deserves to be emphasized that the interest in this study is on item difficulty, not on the responses of particular individuals who took a particular test.

---

[2]In the last five years the sampling of examinees used to calculate the item statistics has been restricted to just juniors and seniors taking the SAT. Furthermore, the sample has been increased to approximately 2,500 for each item rather than 1,500.

Different ability levels were also used as dependent variables. With respect to total verbal performance, the lowest ability level is defined as the lowest scoring 20% of the random sample of 1500 examinees who take a particular test form. The highest ability group represents the top 20% of this random sample of examinees based on total verbal score. The item statistics card report only percent passing each item for each ability level; hence, these percent pass scores must be converted into z-scores prior to analysis. Even though an equivalent equating score (similar to equated delta above) is not available for percent pass at each ability level, and so the particular items are not strictly equated across test forms, we have decided to combine the data across test forms in order to gain some insight about how different ability groups appear to respond to different types of items.

Dependent Variables

v5 Item equated Delta
v88 Percent low ability examinees passing item
(v88 through v94 are used only for correlation tables in Appendix B)
v91 Percent 2nd lowest ability examinees passing item
v92 Percent middle ability examinees passing item
v93 Percent 2nd highest ability examinees passing item
v94 Percent highest ability examinees passing item
v388 z-score of v88 (this variable was used in all regression analyses for low ability examinees)
v394 z-score of v94 (this variable was used in all regression analyses for high ability examinees).

In scoring items, the structure and content of item stems, correct options and incorrect options were recorded using the 19 item variables listed above. A related set of variables were scored for capturing the passage information, but included additional variables which were unique to the text structure--see text variables listed above. In all there are 37 text variables. Also there are 9 text by item variables that apply to location of main idea information. (Four variables--v70,v73,v76,v79--were not included in any of the analyses below because the number of observations per variable were fewer than 3 out of 110 main idea items).

Results and Discussion

Item, Text and Text by Item Predictors of Main Idea Items: Correlation results

In Table 1, which focuses upon main idea items, we see that a number of the item, text and text associated variables are significantly correlated with the equated delta and/or with the z-scores for the low and high ability groups. In Table 1 there are 19 variables that are significant for all three dependent variables (delta, low ability z-score and high ability z-score). We will confine our comments to just these 19 variables. [See Appendix B for means and standard deviations of all independent variables and the correlation of each independent variable with equated delta and each ability level.]

Insert Table 1 about here

There is only one significant item variable (v75) for all three dependent variables: v75 indicates whether the correct answer contains a negation (e.g., "no" "not" "useless" "unconscious") or not. The presence of negations makes the item harder.

Most of the significant correlations relate to the text variables (two variables, v338 & v342 represent text by item interactions). We now present a brief discussion of the remaining 18 variables [the reader should note that many of these significant variables are intercorrelated, therefore they do not all represent independently significant results (this issue is taken up later when we report our regression results)]:

1. The concreteness (v44) of the text contributes the most to making a main idea item easy. Conversely, an abstract text makes a main idea item very hard. This is important to both high and low ability levels. The facilitating effect of concreteness may be due to the availability of a second storage mechanism (a visual one associated with the high imagery of concrete passages--see Paivio, 1971). When verbal storage capacity is overloaded, the visual one may make supplementary space available; if so, this should increase the accuracy of representation and thereby improve performance on main idea items--also see Just and Carpenter (1987) for performance decrement when language capacity is overloaded. The difficulty of abstract text was predicted by our earlier review and occurs as category i under Hypothesis 1.

Table 1
Correlations of Significant Item, Text and Text Associated Variables with Equated Delta and z-scores for Percent Pass for High and Low Ability Examinees for 110 Main Idea Items

Variable Name	Equated Delta [a]	% pass low ability [b]	% pass high ability [b]
v3 Position correct answer	-.19**	-.16	-.13
v4 Coherence of text	.24**	.18*	.24**
v12 Number text words	-.15	-.07	-.19**
v14 Number words in stem	-.18*	-.19**	-.15
v18 Number words in first paragraph	-.22**	-.22**	-.15+
v19 Number words in longest paragraph	-.27**	-.28**	-.22**
v35 Argumentative text	-.39**	-.47**	-.31**
v40 Main Idea information in first sentence, paragraph two	.17	.05	.28**
v42 Number sentences in 1st paragraph	-.20**	-.22**	-.17++
v43 Number sentences, longest paragraph	-.28**	-.33**	-.26**
v44 Concreteness of text	.54**	.55**	.50**
v45 Rhetorical structure: list/describe	.28**	.24**	.29**
v55 Frequency of text's deferred foci	-.20**	-.16++	-.23*
v56 Frequency of text questions	-.28**	-.25**	-.28**
v58 Text is a science excerpt	.37**	.38**	.29**
v59 Text is "about science"	-.20**	-.22**	-.15
v64 Text's frequency of referentials across independent clauses	-.25**	-.21**	-.32**
v66 Sum of three referential codes	-.24**	-.17++	-.30**
v67 Number of text negations	-.35**	-.28**	-.41**
v71 Stem, use of fronting	-.16++	-.12	-.12
v75 Number negations in correct option	-.26**	-.19**	-.24**
v90 Average number words per paragraph	-.19**	-.23**	-.12
v100 Text involves natural science	.32**	.33**	.26**
v337 Rhetorical organization: compare-adversative	-.38**	-.45**	-.22**
v338 Main idea information in middle of text	-.22**	-.24**	-.23**
v342 Main idea information in 1st and/or 2nd sentence and/or 1st short text paragraph	.25**	.28**	.18*

[a] A negative delta correlation makes main idea harder; algebraic sign reversed to make it consistent with z-score results; a positive correlation in this table means the presence of the variable makes the main idea easier (equated delta uses the full ability spectrum.

[b]
** significant, $p < .05$, 2-tailed
* marginally significant $p < .06$, 2-tailed
++ significant, $p < .05$, 1-tailed
+ marginally significant, $p < .06$, 1-tailed.

If a variable was not significant for the 2-tailed test but appeared as one of the variables listed under Hypotheses 1-4 where direction was predicted, we applied a 1-tailed test.

Apart from the research of Paivio (1971), it will be useful to consider the significance of abstractness in a more extended discourse theoretic framework. As far as we can determine, while the effect of abstractness is not explicitly predicted by any of the three text representation models discussed by Abelson and Black (1986), it may be taken as indirect support for the content-functional approach that emphasizes the purpose served by a text (essentially, the 'point' the author is trying to make). That is, an author's purpose in writing a technical (generally concrete) versus non-technical (generally abstract) text probably differs in a number of respects. The degree to which this is so might contribute (over and above the imagery aspect) to main idea difficulty. However, a richer theory for interpreting differences in concreteness would flow from a cognitive interpretation of the sociolinguistic perspective (see Freedle & Duran, 1979). Sociolinguistics emphasizes how different contents can serve different cultural ends as a function of differences in such factors as formality, setting, goal, participants, topic, mode of presentation (written or spoken), etc. (also see Hymes, 1962; Ervin-Tripp, 1964). To give one example, consider the style of technical writing that favors more affirmative and shorter sentences. To help explain such stylistic differences, one notes that some of the purposes served by science are clarity of definition and brevity; this is not always true of the purposes served by nontechnical prose (e.g., the humanities). That is, since the style of presentation is in part a reflection of its underlying social purpose, this might help account not only for stylistic differences across different content areas but would also help to explain why these differences exist.

2. An argumentative (v35) text (representing a special point of view adopted by the author) is next most important; an argumentative text makes main idea items quite difficult compared with nonargumentative texts. Low ability students are very strongly affected by argumentative texts. The fact that this variable is significant (quite apart from ability level) can be taken as weak support for the discourse theoretical frameworks developed by Meyer (1985) and Grimes (1975) who investigated the top-level rhetorical representation of texts (also see next paragraph). This result was anticipated by category c under Hypothesis 1.

3. This is a special type of Grimes compare, called compare-adversative (v337) where one component is stated to be superior (see Meyer, 1985); this form makes main ideas more difficult. The significance of this variable provides stronger theoretical support for the top-level discourse representation scheme of Meyer (1985) and Grimes (1975). Cognitively, main ideas are harder with this type of text organization because several concepts are being contrasted (and one of these is being favored by the author). Other types of

Grimes organizers, by contrast, appear to make main idea items easier, especially List type structures (see v45 under #8 below) which presents a collection of related ideas. The significance of v337 was anticipated under category c under Hypothesis 1.

4. As the number of text negations (v67) increases, the main idea item becomes harder. The importance of negations on language processing has been stressed by Carpenter and Just (1975)--also see Just and Carpenter (1987). The significance of this variable was anticipated by category a under Hypothesis 1.

5. If the passage is an excerpt of science (v58), the main idea item is easier. Incidentally, the somewhat lower significance of 'about science' (v59), which makes items more difficult, may be considered weak support for the content-functional model advanced by Abelson and Black (1986). That is, since the content of the exposition makes a difference (presumably because it alters the author's stated and implied purposes in writing the text) it becomes a relevant consideration in evaluating the 'point' of a text representation scheme. The relevance of a sociolinguistic (see the paragraph on concreteness, v44, above) perspective is again pertinent here. This variable does not fall under the categories listed for Hypothesis 1.

6. If the passage consists of natural science (v100) material (it can be either an excerpt of science or 'about science'), it makes the main idea item easier. Since most natural science texts are "excerpts of science" the conclusion reached for variable v58 above still holds. The significance of this category was not anticipated by Hypothesis 1.

7. As the number of sentences in the longest paragraph (v43) increases, the main idea item gets harder. This implies that the more information there is, the more difficult it is to decide what the main idea is. Category g under Hypothesis 1 anticipated the significance of this type of variable.

8. The Grimes structure called List and/or Describe (v45) makes main idea items easier. A list (and/or a describe) structure is basically a series of statements about members of a category; often there is no intrinsic ordering to the members of the list. This is the second variable supporting the Meyer (1985) and Grimes (1975) coding scheme for top-level text information. This variable falls under category c of Hypothesis 1.

9. As the number of questions posed in the text (v56) increases, the main idea item gets harder. This may relate to the uncertainty about what the author is asserting. That is, the more questions asked, the less clear it may

be about what the author is really asserting. It is not obvious that the significance of this variable would be predicted by any of the text representation models cited above. Hypothesis 1 did not anticipate the significance of this variable.

10. The more words in the longest paragraph (v19) the harder the main idea item. This suggests that as the amount of material increases, the examinee has to work harder to determine what the central idea is. (Obviously v19 is correlated with v43 described above.) This again implicates category g under Hypothesis 1.

11. When the overall "coherence" (v4) of the passage is high (meaning the same concepts of the opening sentence appear throughout all the paragraphs including the final sentences), the main idea is easier to locate. Presumably, this implies that when only a few concepts are used throughout the text, it is easier to decide what the main idea is, either due to repetition effects and/or because only a few concepts are being discussed. Although it has not been directly tested, it seems likely that high coherence as measured here would be consistent with Kintsch's (1974) representation scheme since there would be less depth to a highly coherent passage (with a few arguments repeated over and over) in comparison with a passage of similar length which was low in coherence (implying many new arguments with fewer repetitions). The significance of this variable was not anticipated by Hypothesis 1 primarily because we did not carry out an exact Kintsch-type (1974) scoring.

12. If the main idea is mentioned in the first and/or second sentence of the text and/or in the first short paragraph (v342), this makes main idea items easy to get correct. This suggests that when the main idea is in a position where it is normally expected to be--near the opening of the passage--this makes it easier to confirm that it is in fact the main idea. This result lends support for Kintsch's (1974) early propositional coding model. Early sentences are typically higher in the hierarchy of propositions than later sentences, hence they should be easier to retrieve relevant information from. This result also appears to support our generalization of Kieras' (1985) findings. Also, without further analysis, it would appear that Kieras' (1985) findings generalize to multiparagraph texts and to nontechnical prose as well. However for another view of this idea see Appendix C. The significance of this variable was anticipated by Hypothesis 4.

13. As the number of words in the first (v18) paragraph increases (see also v42 below), the main idea items become harder. This suggests that the opening paragraph is expected to contain the main idea--whether that is true

or not; so, as the first paragraph grows in length, examinees find it more difficult to decide whether the main idea is or is not present in the opening paragraph. The significance of this type of category was anticipated by category g of Hypothesis 1.

14. As the number of deferred foci (v55) increases in the text, this makes the items harder. Deferred foci delay the introduction of the semantic substance of a sentence by introducing a 'dummy' subject (e.g., "It is the case that things are difficult"). This delay might introduce additional uncertainty regarding whether these sentences are asserting clear main idea information or not. That is, this and other types of "frontings" (Freedle, Fine, & Fellbaum, 1981) can be thought of as qualifying or altering the impact of the sentence subject and hence adding cognitive complexity to what the main thrust of the sentence is. The significance of this type variable was anticipated by category d of Hypothesis 1.

15. The more pronoun referential expressions (v66) that are in the text the harder the main idea items. If many referential expressions are present per sentence, this increases the amount of "bridging" that must be accomplished (see Clark & Haviland, 1977) in order to determine what the sentence is asserting. If such sentences contain the main idea (or allude to it), having many referential expressions should interfere with determining a clear statement of the main idea. The significance of this type variable was anticipated by category b of Hypothesis 1.

16. When the main idea information is located in the middle of the text (v338) this makes it harder to get main idea items correct. This finding is probably due to the fact that examinees expect the main idea to be located at the beginning (or end) of the passage, not in the middle. Both ability groups are about equally sensitive to this variable. This result provides some support for Kintsch's (1975) early propositional representation of text information and supports our extension of Kieras' (1985) results as well. Middle propositions are probably embedded deeper in the text than earlier propositions. Hence they should be harder to retrieve relevant information from regarding main ideas. The significance of this type variable was anticipated by Hypothesis 4.

17. Variable v64 indicates that as the number of referentials across independent clauses (as opposed to primarily within clauses) in the text increases, the main idea item becomes increasingly difficult. This result was anticipated by category b of Hypothesis 1.

18. Variable v42 indicates that as the number of sentences in the first paragraph increases, this makes main idea items harder. This type variable falls under category g of Hypothesis 1.

Based on just the zero-order correlations presented in Table 1, what can we conclude concerning Hypothesis 1 (which states that the nine categories found to affect comprehension difficulty in the experimental literature will also affect comprehension difficulty as measured by multiple-choice tests)?

From the results in Table 1 we see that the following six categories are confirmed as influencing multiple-choice comprehension difficulty for main idea reading item: negations, referentials, rhetorical organizers, fronted structures, paragraph length and abstractness of text. There seems to be substantial evidence that Hypothesis 1 is supported for six of the nine categories. Therefore it appears that responses to multiple-choice reading tests are not that different from responses to comprehension materials presented in controlled laboratory studies. This result therefore casts some doubt on some assertions made recently by Royer (1990) and Katz et al. (1990) which asserted that multiple-choice comprehension tests do not measure comprehension but only a generalized reasoning ability.

Regarding Hypothesis 4 which predicts a significant effect due to the relative location of main idea information we see that v342 & v338 are significant. Therefore Hypothesis 4 seems to be fully confirmed by just the correlational results.

Related correlational findings taken from the GRE multiple-choice reading items will be presented later in this report; in general we shall see that the GRE results further confirm many of the findings reported above for SAT correlations.

Examination of the full table of intercorrelations (not presented here) indicates that many of the above significant variables are closely interrelated (e.g., natural sciences often are classified as having concrete passages, and furthermore the science passages contain fewer text negations than, say, the humanities passages do, etc.). In order to determine which of these variables contribute independent variance to the prediction of equated delta we need to use other statistical techniques. To achieve this we present below several regression analyses of main idea difficulty.

Criteria for admitting variables into the stepwise regressions

For all stepwise regressions the following criteria were used for admitting variables into the final solution. All variables were available for possible selection. Each new variable that was admitted into the solution had to yield a significant individual F value, and, in addition, the new F values for all previously admitted variables had to be significant. If the next variable admitted showed a nonsignificant F, then the previous solution was considered the final one.

Companion regression analyses were also run where only the variables that significantly correlated with one or more of the three dependent variables (see Table 1) were considered for use as predictor variables; otherwise the same criteria just mentioned applied to these companion analyses. This alternative way to select possible predictor variables represents one way to restrict the large number of predictor variables in our study.

Stepwise regression analysis of main idea items: Equated Delta as the Criterion

In Table 2 we present the regression results for predicting the equated delta values of 110 main idea items. First we note that the 8 significant independent predictors account for 58% of the item difficulty variance.[3]

Insert Table 2 about here

Implications for Hypotheses 2 & 4 for equated delta: main idea items.

Hypothesis 2 says that the nine categories listed in Hypothesis 1 should provide independent predictive information concerning main idea difficulty. Of the nine, Table 2 reveals that four are seen to be significant independent predictors of main idea difficulty. They are: concreteness (v44), paragraph length (v19), rhetorical organization (v337), and frontings (v55). We see

---

[3]We realize that using a large number of predictor variables can capitalize on chance, making some particular variable seem more important than it in fact might be if the study were replicated with another 110 passages and their associated items. However, we discuss individual variables here to give the reader a flavor of how to interpret the scored variables which happen to yield the strongest correlations wit the criterion. What we do not expect to change in any replication are the general categories into which the significant variables fall.

Table 2

Stepwise Regression Analysis
Predicting 110 Main Idea Delta Values

	Variable	F value[a,b] of each predictor	Percent Variance	Zero-order Correlation	Source
v44	Concreteness of text	35.0	30%	.54	text
v19	Number words in longest paragraph	8.5	06%	-.27	text
v337	Compare-argumentative text	14.2	05%	-.38	text
v3	Position of correct option	13.6	05%	-.19	item
v4	Coherence of text	10.8	04%	.24	text
v55	Number of clefts in text	6.3	03%	-.20	text
v342	Main idea 1st and/or 2nd sentence or first short paragraph	9.6	03%	.25	text by item
v40	First sentence of 2nd paragraph	6.6	03%	.17	text by item

a

The variables are listed in the order they were extracted by the regression routine. The algebraic sign of the zero-order correlation has been reversed so as to agree with the convention adopted in the other tables of this report. A positive correlation means that the variable facilitates getting the item correct.

b

The overall $F(8,101) = 17.7$, $p < .01$. The multiple R taken from the final solution is .764; the R squared is .584. The individual F values for each predictor are taken from the final regression step. Individual F values are significant at $p = .05$ when they equal 3.94 or larger; they are significant at $p = .01$ when they equal 6.88 or larger.

that these results provide modest support for Hypothesis 2 as they apply to one reading item type: main idea items. For a related set of findings (see Freedle & Kostin, 1991) it was found that GRE main idea items yield only two of the above categories (paragraph length and frontings) as providing independent category information for main idea difficulty. [However, the relatively small GRE sample size for main idea items may have attenuated the possible significance of other categories.]

Two of the remaining four independent predictors listed in Table 2 apply to Hypothesis 4 (v342 & v40). This result indicates that main idea information located early in the text makes these main idea item easy. Hence half of Hypothesis 4 has been confirmed by this particular result. Hypothesis 4 was also supported in our analyses of GRE main idea reading data (Freedle & Kostin, 1991).

The companion regression (which admitted only predictors having a significant correlation with item difficulty) yielded the identical set of predictor variables just described.

Stepwise regression analyses of main idea items for low and high ability examinees

Now we consider the separate analyses for the performance of high versus low ability examinees. In Table 3 we have added the predictors for equated delta (taken from Table 2) to facilitate comparisons with the predictors found for high and low ability examinees. For low ability there are eight significant and independent predictors of item difficulty accounting for 59% of the item difficulty variance, while for high ability there are six significant predictors which account for 46% of the variance.

Insert Table 3 about here

We see that there are 10 different predictors that account for one or both ability level groups: v44, v3, v4, v342, v337, v55, v40, v43, v89, v67. The first four of these are independent predictors for both the high and low groups. [These four were also independent predictors for equated delta.] The remaining variables differ as to which group they aid in predicting main idea responses. These different variables may reflect possibly different strategies that the two groups are using in answering main idea items.

Table 3

Stepwise Regression Results for 110 Main Idea Items for
Three Criterion Variables:
Equated Delta and Two Ability Groups

Predictor Variable	Equated Delta F value	Equated Delta Percent Variance	Low Ability F Value	Low Ability Percent Variance	High ability F value	High ability Percent Variance
	a b		b		b	
v44	35.0	30%	33.2	30%	24.5	25%
v19	8.5	06%	--	--	--	--
v337	14.2	05%	21.5	08%	--	--
v3	13.6	05%	8.5	04%	6.0	03%
v4	10.8	04%	4.4	02%	6.6	04%
v55	6.3	03%	4.5	02%	--	--
v342	9.6	03%	8.4	04%	5.8	03%
v40	6.6	03%	--		16.4	06%
v43	--		18.0	08%	--	--
v89	--		4.8	02%	--	--
v67	--		--		8.8	07%

Overall Predictability for each of three criteria:

Eq. Delta	Low Ability	High Ability
F(8,101)=17.7**	F(8,101)=18.3**	F(6,103)=14.8**
Mult.R .76	Mult.R .77	Mult.R .68
R Sq. .58	R Sq. .59	R Sq. .46

** = overall F value significant, $p < .01$.

---

a

The first eight predictor variables are listed in the order they were extracted with equated delta as the criterion:

Table 3 (cont.d)

v44 = text variable; abstractness/concreteness
v19 = text variable: number of words in longest paragraph
v337 = text variable: comparative-alternative rhetorical type
v3 = item variable; ordinal position of correct answer
v4 = text variable; lexical coherence over paragraphs
v55 = text variable: frequency of deferred foci
v342 = text by item variable: main idea in 1st or 2nd sentence and/or in first short paragraph
v40 = text by item variable: main idea information in 1st sentence paragraph two.
v43 = text variable: number sentences in longest paragraph
v89 = text variable: average number words per sentence
v67 = text variable: number of negations

b
Each individual F value listed for each predictor variable is significant at $p < .05$ or beyond.

If we focus only on those differences which are more easily interpreted we have the following. The low ability show considerable difficulty in interpreting the main idea of compare-argue passages (v337). Their zero-order correlation was -.45 . The high ability examinees also show difficulty with this structure (r= -.22), but it does not figure as an independent predictor of their main idea difficulty. The result suggests that many of the low ability examinees may not fully appreciate the meaning of comparative-argue passages; hence the type of rhetorical organization of a passage appears to make a difference across ability groups. The other variable that is relatively easy to interpret is v43 (number of sentences in the longest paragraph). The longer the passage paragraphs are the more difficult it is for low ability people to find the main idea (r = -.33). High ability people also have some trouble with long paragraphs (r = -.26) but this variable fails to yield independently important information in predicting overall main idea difficulty for them.

Two variables are more important for the high ability students: number of text negations (v67) and occurrence of the main idea in the first sentence of the second paragraph (v40). If the high ability appear to pay more attention to text negations than the low, this might help account for the larger negative correlation they have (r= -.41) than the low ability people (r = -.28). Negations are of course important because they alter the truth value of text assertions; high ability people may be very sensitive to text elements that can potentially alter the truth value of what they are reading. The second variable that high ability people differ on is whether the topic occurs in the first sentence of the second paragraph. High ability people are facilitated in finding the main idea if it occurs in this text position (r = .28) while low ability presumably do not specifically scan this part of the passage in looking for the main idea (r = .05).

. Ability level regression results and its implications for Hypotheses 2 & 4: Main idea items.

The following four categories (taken from Hypothesis 1) provide independent predictive information for low ability examinees: abstractness (v44), frontings (clefts, v55), paragraph length (number of sentences in longest paragraph, (v43), and sentence length (average number of words per sentence (v89). Hence there is modest support for Hypothesis 2 using the low ability examinee's results. Also the early location of main idea information (v342) facilitates low ability performance; this confirms half of Hypothesis 4 for low ability examinees. Incidentally, Anderson & Davison (1988) also discuss the fact that lower ability 7th graders experience greater difficulty with longer sentences than high ability students.

The following two categories (taken from Hypothesis 1) provide independent predictive information for high ability examinees: abstractness (v44), text negations (v67). Thus the results for high ability people provides rather poor confirmation of Hypothesis 2 for main idea items. Also the independent significance of early location of main idea information (v342 and v40) provides support for half of Hypothesis 4 for high ability people.

It therefore appears that low ability people provide better support for Hypothesis 2 than high ability people; both groups provide similar support for half of Hypothesis 4.

Companion regression analyses for high and low ability groups were also run using just the significant correlated variables as predictors. For high ability, the regression result is identical to that already reported above. For low ability, the regression result is basically similar except that there are only seven significant predictors (instead of eight) which account for 57% of the variance. (The missing variable is v89).

Predicting the full item sample (n=285 items) using the set of predictor variables developed for main idea items: Correlations

In Table 4 we present the significant correlations of each variable with the full item sample. While these variables were intended primarily to reflect main idea difficulty, they nevertheless appear to do a fair job describing most of the reading items used in the SAT reading section (75% of the item types which occur in the SAT reading section consist of the three types studied here: main ideas, inferences, and explicit statement items.)

Insert Table 4 about here

We quickly compare what is different across Tables 1 and 4 prior to conducting our regression analyses. Six new delta variables appear here which were not significantly correlated with main idea items (see Table 1 above). These new variables are: v14, v60, v61, v68, v78, v96. Ten variables which were significant for main idea items (see Table 1) are no longer significant for the full item sample: v3, v4, v42, v45, v55, v56, v59, v64, v66, v338 are no longer significant for the full item sample. The presence of some of these new variables (v60 and v61) in Table 4 (but not Table 1) is easy to explain: v60 represents a code for whether the item is a main idea item or not, v61 represents a code for whether the item is an inference item or not. (Table 1

Table 4

Correlation of each Significant Variable with 285 Reading Items consisting of Main Ideas, Inferences and Explicit Statements

		Three Criterion Variables		
Variable	Brief Description	Equated Delta[a]	% pass z-score Low Ability	% pass z-score High Ability
v14	Number words in item stem	-.21**	-.23**	-.18**
v15	Number words in correct option	-.10	-.06	-.13**
v18	Number words in 1st paragraph	-.13**	-.15**	-.12**
v19	Number words in longest paragraph	-.13**	-.17**	-.13**
v35	Argue	-.15**	-.23**	-.13**
v40	Main idea information in 1st sentence, second paragraph	.14**	.10	.17**
v43	Number sentences in longest paragraph	-.08	-.12**	-.08
v44	Concreteness	.35**	.35**	.33**
v58	Science excerpt	.18**	.20**	.12**
v60	Main idea items	.18**	.19**	.12**
v61	Inference items	-.25**	-.23**	-.25**
v62	Explicit Statement items	.07	.04	.13**
v67	Text negations	-.15**	-.14**	-.15**
v68	Stem, use of hedge	-.13**	-.15**	-.13**
v75	Negatives in correct option	-.16**	-.16**	-.15**
v78	Negatives in incorrect options	-.21**	-.18**	-.23**
v89	Average number words in sentence	-.10++	-.11++	-.05
v90	Average number words/paragraph	-.12**	-.14**	-.10++
v96	Average sentence length in first paragraph	-.13**	-.13**	-.07
v100	Natural science content	.17**	.17**	.12**
v337	Compare-adversative	-.20**	-.24**	-.12**
v342	Main idea is in 1st or 2nd sentence and/or first short paragraph	.15**	.15**	.08

[a] The delta algebraic sign has been reversed for ease of comparison with the z-scores. All positive correlations are interpreted as facilitating getting an item correct. If a correlation was significant for any (or all) of the criterion variables, it was included in this table.

[b] ** significant, $p < .05$, 2-tailed; ++ significant, $p < .05$, 1-tailed. If a variable was not significant for the 2-tailed test but appeared as one of the variables listed under hypothesis 1-4 where direction was predicted, we applied a 1-tailed test.

represented only main idea items.) We see that the positive (delta) correlation for main ideas and the negative one for inference items indicates that main ideas are easier than inference items.

Hypothesis 1 lists nine categories (see above). Table 4 implicates the following 5 categories for the full item set (n=295): paragraph length (v18, v19, v43 & v90), concreteness (v44), rhetorical organization (v35 & v337), negations (v67 & v75), sentence length (v89 & v96). Also Hypothesis 4 which deals with the location of main idea information is partly supported (v342 & v40).

Since all these variables are to some degree intercorrelated, we again need to use a regression analysis to determine which of these variables provides independent prediction of the variance of all 285 reading items.

Stepwise regression results for analyzing the full item sample (all three item types) for each of three criterion variables: equated delta, and high and low ability levels

Table 5 provides the relevant results from the stepwise regressions predicting item difficulty for the full item sample (n=285 items). We see that 29% of the item difficulty variance can be accounted for by 9 variables.[4] These variables relate to the categories of Hypothesis 1 in the following way: v44 (concreteness), v78 (incorrect option negations), v337 & v35 (rhetorical organization: compare-adversative), and v19 (paragraph length: number words in longest paragraph). Thus four of the nine categories of Hypothesis 1 are supported by the full item sample; this provides modest support for Hypothesis 1. The fact that these same variables provide independent predictability provides modest support for Hypothesis 2. Hypothesis 4 concerning effects of differing locations of main idea information, while not specifically predicted for the full item sample, nevertheless appears to be

---

[4]The significant stepwise regression using the full set of items could be argued to be due solely to the influence of the main idea items. To check this possibility we separately analyzed the sample of inferences and explicit statement items (n=175) by the stepwise procedure. This yielded 8 significant and independent predictors for the delta criterion (v44, v61, v78, v96, v40, v41, v16, v65) for a total $F(8,166) = 7.3$, $p < .01$. This accounted for 26% of the variance. For the low ability group five variables were significant (v44, v96, v61, v78, v40) for a total $F(5,169) = 7.9$, $p < .01$. This accounted for 19% of the variance. For the high ability group seven variables were significant (v62, v78, v44, v56, v40, v41, v15) for a total $F(7,167) = 8.4$, $p < .01$. This accounted for 26% of the variance. Clearly the significance of the full item set (n=285) is not solely due to the presence of the main idea items in the sample.

valid here given the significance of variables v342 and v40. The significance of variable v61 simply indicates that the inference items differ significantly in overall difficulty level from the remaining two item types; by itself this does not apply to any of our hypotheses.

High ability examinees accounted for 25% of the variance for the full item sample. The following categories of Hypothesis 1 are validated by the high ability group: concreteness (v44), negations (v78) and paragraph length (v19). For the low ability group concreteness (v44), paragraph length (v19), rhetorical organization (v337) and vocabulary (v17) are the categories supported. The low ability group apparently provides about the same level of support for Hypothesis 1 as the high ability group. Furthermore since these category results each provide independent predictability this indicates that Hypothesis 2 also is modestly supported by the two ability groups. Incidentally, this is the first analysis for which vocabulary yielded a significant result. Just and Carpenter (1987) indicate that vocabulary seems to be a more critical variable in predicting low as compared with high ability reading comprehension performance. Hence this particular result seems consistent with the Just and Carpenter (1987, p.460) finding.

Insert Table 5 about here

Companion stepwise regressions were also run using only significantly correlated variables as predictors (see significant variables in Table 4). For equated delta and high ability examinees as the criteria, the two stepwise regressions are identical to those reported in Table 5. For low ability examinees the results are similar but not identical to that presented above. The following variables were significant: v44, v61, v19, v337, v78 which accounted for 24% of the variance [$F(5,279) = 17.6$, $p < .01$]. Variable v17 dropped out of this analysis, but variable v78 is now added to the current regression.

Hierarchical regressions

Main Idea items: hierarchical regressions

Methodologists indicate that a hierarchical regression analysis is called for when comparing the relative contribution of two sets of variables (the two sets, for example, being item variables and all remaining text and text associated variables). A test of Hypotheses 3a-d necessarily involves a contrast of the effects of all item variables versus combinations of the

Table 5

Stepwise regression of 285 items for three criterion variables: Equated delta, percent passing for low and high ability groups[a]

Predictor Variable[e]	Eq.Delta F value[b,c]	Low ability F value[d]	High ability F value[d]	Source
v44	33.7 13%	31.9 12%	32.4 10%	text
v61	20.9 06%	19.5 05%	18.5 06%	item
v78	7.1 03%	--	9.8 03%	item
v19	6.4 02%	13.4 03%	5.0 01%	text
v337	10.5 01%	9.4 02%	--	text
v40	8.0 01%	--	6.9 02%	text by item
v342	6.4 01%	--	--	text by item
v35	6.0 01%	--	--	text
v15	5.5 01%	--	5.5 01%	item
v17	--	6.0 02%	--	text

a The sample of 285 reading items includes three types of items: main ideas, inference, and explicit statement items. The individual F values for each variable are significant beyond $p < .05$. b

The overall F value for each of the three criterion variables is as follows:

Eq. Delta	Low Ability	High Ability
F(9,275)=12.7**	F(5,279)=17.8**	F(6,278)=15.1**
Mult.R .54	Mult.R .49	Mult.R .50
RSq. .29	R Sq. .24	R Sq. .25

c The individual F value listed for each predictor variable is significant at $p < .05$ or beyond.

d Each of the percent pass scores was converted to a z-score prior to the regression runs.

e v44 = concreteness of text; v61 = inference item type; v78 incorrect option use of negation; v19 = number words in longest paragraph; v337 = rhetorical organization: compare-adversative; v40 = main idea information in 1st sentence, 2nd paragraph; v342 = main idea information in 1st and/or 2nd sentence and/or first short paragraph of text; v35 = argumentative text; v15 = number words in correct option; v17 = number words with three or more syllables in 1st 100 passage words.

remaining variables (the text and text associated variables). Hence a series of hierarchical regression analyses was used to evaluate Hypotheses 3a-d.

Table 6 presents the results relevant for Hypothesis 3b which states that item variables alone will account for the significant predictability of item difficulty while text and text associated variables when added after extracting item effects, will not be significant. [Remember, this hypothesis is derived from assertions made by Royer (1990) and Katz et al. (1990)].

Insert Table 6 about here

Table 6 shows us that, for main idea items, when all item predictors are extracted as the first set of variables they account for a significant amount of item difficulty variance (20.7%) only when all the ability groups are used [$F(12.97) = 2.1$, $p < .05$]. For high and low ability groups, neither result shows that item variables alone account for significant variance: for high ability 15.5% of the variance of item difficulty is accounted for, which is not significant [$F(12.97) = 1.5$, $p > .2$]], while for low ability 15.2% variance is accounted for; this is also not significant [$F(12.97)=1.4$, $p > .2$].

However, the same Table 6 also shows us that when text and text associated variables are added as the second set of predictors, the additional variance accounted for is significant for high and low ability groups as well as for all ability groups ($p < .01$) in all cases--see Cohen & Cohen, p. 146-147). In particular, in excess of 45% of the item difficulty variance is accounted for by text and text associated predictors. This represents approximately three times as much variance as that accounted for by the item predictors.

This set of results tells us several things: (1) Hypothesis 3b is not supported for main idea items; hence the claims made by Royer (1990) and Katz et al. (1990) appear to be incorrect; and, (2) Hypothesis 3d is not supported because item variables in fact do <u>not</u> account for <u>more</u> variance that do the remaining predictor variables; in fact the variance accounted for by item predictors is not only much lower than for the remaining variables, but is in some cases not even significant.

Now we will evaluate Hypotheses 3a and 3c which deal with the contrast between item variables and just the text variables (i.e., the text associated variables are left out of the computations) as it applies to main idea items.

Table 6

Hierarchical Regression of 110 SAT Main Idea Reading Items: an Evaluation of Hypotheses 3a, 3b, 3c and 3d

Type of Items	Percent Variance	F value	p level
All ability levels			
1st set (items)	20.7 %	F(12,97) = 2.1	.05
2nd set (T+T*I)[a]	51.8%	F(36,61) = 3.2	.01
All predictors	72.5%	F(48,61) = 3.4	.01
High ability group			
1st set (items)	15.5%	F(12,97) = 1.5	n.s.
2nd set (T+T*I)	45.6	F(36,61) = 2.0	.01
All predictors	61.0%	F(48,61) = 2.0	.01
Low ability group			
1st set (items)	15.2	F(12,97) = 1.4	n.s.
2nd set (T+T*I)	57.0%	F(36,61) = 3.5	.01
All predictors	72.2%	F(48,61) = 3.4	.01
All ability groups			
1st set (items)	20.7%	F(12,97) = 2.1	.05
2nd set (text only)	46.1%	F(31,66) = 3.0	.01
All predictors	66.8%	F(43,66) = 3.1	.01
High ability group			
1st set (items)	15.5%	F(12,97) = 1.5	n.s.
2nd set (text only)	40.1%	F(31,66) = 1.9	.05
All predictors	55.6%	F(43,66) = 1.9	.01

Table 6 (cont.d)

Low ability group			
1st set (items)	15.2%	F(12,97) = 1.4	n.s.
2nd set (text only)	52.7%	F(31,66) = 3.4	.01
All predictors	67.9%	F(43,66) = 3.2	.01

[a] The symbol T+T*i means all the text predictors (T) plus the text by item predictors (T*i).

The last half of Table 6 provides the relevant results. Since the same item variables are extracted we get the identical results as found for the first half of Table 6: as before, the item variables alone only account significantly for the data dealing with all ability groups; high and low ability groups by themselves do not show a significant item effect (for the block of item predictors). Text variables on the other hand do account for a substantial proportion of item difficulty variance. In particular after the item variables are extracted, text variables account for 46.1% of the variance for all ability groups, and for 40.1% and 52.7% of the variance for high and low ability groups, respectively. Therefore Hypotheses 3a and 3c are not supported: it is clear that text variables alone are superior predictors of reading item difficulty while item variables play a very minor role for main idea items.

A broader interpretation of these findings will be presented below after we examine the hierarchical regression results using the full item (n=285) sample. Table 7 presents the relevant results for the full item sample.

Insert Table 7 about here

The full item sample (n=285): hierarchical regressions

The full item set shows that item variables now play a significant role for high and low ability groups as well as for all ability groups combined. The percentage accounted for is relatively low--12.6% to 13.8%--but it is significant ($p < .01$, in all cases). Hypotheses 3b and 3d are nevertheless not supported because we see in Table 7 that, after the item variables are extracted, text and text associated variables also account for a significant proportion of the item difficulty variance--from 19.1% to 25.3%, significant at $p < .01$ in every case. It is clear that item and text plus text associated variables play about an equal role in determining reading item difficulty. Yet Hypotheses 3b and 3d are still not supported because these hypotheses maintain that either item variables alone account for all significant effects (Hypothesis 3b) or that item variables play a dominant role with respect to other predictor variables in predicting item difficulty. Inspection of the last half of Table 7 shows that the same conclusion applies to the evaluation of Hypotheses 3a and 3c for the full item sample.

Table 7

Hierarchical Regression Analyses of 285 SAT Reading Items: an evaluation of Hypotheses 3a, 3b, 3c, and 3d

All ability groups			
1st set (items)	13.8%	F(12,272) = 3.6	.01
2nd set(T+T*I)[a]	22.4%	F(36,236) = 2.3	.01
All predictors	36.2%	F(48,236) = 2.8	.01
High ability group			
1st set (items)	13.5%	F(12,272) = 3.6	.01
2nd set (T+T*I)	19.1%	F(36,236) = 1.9	.01
All predictors	32.6%	F(48,236) = 2.4	.01
Low ability group			
1st set (items)	12.6%	F(12,272) = 3.2	.01
2nd set (T+T*I)	25.3%	F(36,236) = 2.7	.01
All predictors	37.9%	F(48,236) = 3.0	.01
All ability groups			
1st set (items)	13.8%	F(12,272) = 3.6	.01
2nd set (text only)	17.7%	F(30,242) = 2.1	.01
All predictors	31.5%	F(42,242) = 2.6	.01
High Ability group			
1st set (items)	13.5%	F(12,272) = 3.6	.01
2nd set (text only)	13.6%	F(30,242)= 1.51	.05
All predictors	27.1	F(42,242) 2.1	.01

Table 7 (cont.d)

Low ability group			
1st set (items)	12.6%	F(12,272) = 3.3	.01
2nd set (text only)	21.2%	F(30,242) = 2.6	.01
All predictors	33.8%	F(42,242) = 3.0	.01

[a] The symbol T+T*i means all text predictors (T) plus all text by item interaction predictors (T*i).

An explanation of why Hypotheses 3a to 3d are not supported

As we have already pointed out, the experimental literature has recently maintained that multiple-choice reading tests are not really tests of reading comprehension at all, but instead are merely measures of general reasoning (see especially Royer, 1990). These assertions were formulated in light of the findings that there is an ability for examinees to correctly guess the answer to some multiple-choice reading items above chance level even when the examinees have not read the relevant passage (Royer, 1990; Katz et al., 1990; Tuinman, 1973-1974) and in light of the apparent finding that at least one purported "item" variable was the major predictor of reading item difficulty (see Royer's, 1990, interpretation of Drum et al.'s, 1981, study).

We have just seen that by taking these various assertions at face value, we were led to formulate a hypothesis consisting of four variants, none of which provides an adequate account of our multiple-choice SAT reading data. What type of hypothesis would then account for our current set of findings? In particular, why is it that text and text associated variables do so well in predicting main idea items vis-a-vis item predictors, and why do both item and text variables do about equally well in accounting for the full item sample? We shall now outline some reasons for this pattern of results.

Suppose we grant the finding that some reading items can in fact be correctly guessed at levels greater than chance in the absence of reading the passage (Katz et al., 1990; Tuinman, 1973-1974). This means that we grant that at least part of what a multiple-choice test may be measuring is something called "reasoning" (Royer, 1990). However, if a multiple-choice reading test has a valid comprehension component operating, then making the passage available to examinees should significantly augment the percent correct responses, over and above those achieved by sheer guessing alone. This in fact happens (see Katz et al., 1990). Now, cognitively what does it mean to assert that the reading passage itself exerts a significant effect on multiple-choice item correctness? One way to try to study this question is to ask what are the salient features of the reading passage that are significantly correlated with item difficulty (given that the passage is present). The various text and text associated variables which we have defined in this study are precisely the types of measures that one can use to help to identify what variable aspects of a passage are contributing to comprehension difficulty. Why does this make sense? Here is one rationale.

For the moment let us totally ignore the contribution of the guessing component with regard to item correctness. Suppose many items have been written that turn out to be hard not because the passage is hard to understand

but because each item contains an unfamiliar word. An item predictor that assesses the contribution of vocabulary to item difficulty would presumably show that such items are hard only because of the presence of these unfamiliar words. None of the text predictors would be strongly correlated with such difficult items because, by assumption, it is not the difficulty or ease of the passage that makes these particular items difficult, it is rather a characteristic of the item itself that contributes solely to difficulty. Such items would presumably be caught early in the test construction phase (in assembling a group of items for a multiple-choice reading test) and would be eliminated as obviously irrelevant to the task at hand: to construct items that reflect passage difficulty. Such an idea suggests that most of the items which are finally selected to assess passage comprehension are difficult or easy primarily as a function of passage characteristics with only minor contributions possibly being due to other remaining item characteristics (such as use of negations in the options--see Carpenter & Just, 1975, for discussion of how use of negations affects comprehension). If this is so then it is not surprising to find that text and text associated variables are strongly correlated with reading item difficulty.

Now, why might text and text associated variables account for such a large percent of main idea variance while item variables account for so little? First of all, our variables were chosen specifically to try to capture large and small differences of the total text; this should make such variables better predictors of main idea items, which of course deal with the entire passage, than of other types of items such as explicit statement or inference items which typically deal with only limited portions of the total text. This helps to explain why text and text associated variables account for such a large proportion of main idea variance and why they typically do not account for as much of the variance for the full item sample. Also main idea items were found to have less variability in terms of how the items were structured: for example, they seldom employed negations (in contrast to use of negations in inference and explicit statement items--also see Freedle & Kostin's, 1991, analysis of GRE reading items in this regard). The relative lack of variation in item information of course means that item variables cannot play a major role in predicting item difficulty since there must be at least intrinsic variability in the scoring of a predictor variable before it can possibly function as an effective predictor. But for inference and explicit statement items there was greater variability in the item structure; because of this it is not surprising to have found that the block of item predictors did in fact account for a significant, though small, proportion of the variance.

In conclusion we feel that the demonstration that Hypotheses 3a to 3d are not supported serves as evidence that multiple-choice reading tests (such as the SAT reading items) certainly do function as tests of reading comprehension. They are tests of comprehension because item <u>difficulty</u> has been demonstrated to be a significant function of <u>text</u> and text associated variables not only for main idea items but for the full set of reading items as well.

A comparative analysis of SAT and GRE Main Idea Items: Correlations

Freedle and Kostin (1991) analyzed GRE main idea reading items using an identical set of predictor variables as have been used in our current study. It will be useful to examine which variables proved to be significantly correlated for each data set in order to gain some insight into the stability of our predictors. Of the 22 variables listed in Table 1 above that correlate significantly with equated delta ($p < .05$, 2-tailed) [i.e., v3, v4, v18, v19, v35, v42, v43, v44, v45, v55, v56, v58, v59, v64, v66, v67, v75, v90, v100, v337, v338, v342] the following 10 variables also proved to be significantly correlated with equated delta using the GRE main idea sample [all ten of the variables listed below were significantly correlated with all three of the SAT criterion variables--see Table 1]:

v4 (coherence)
v19 (number of words in text's longest paragraph)
v35 (author of text takes an argumentative stance)
v55 (frequency of clefts)
v56 (frequency of text questions)
v64 (frequency of pronoun referentials across independent text clauses)
v66 (sum of all text pronoun referentials)
v67 (frequency of text negations)
v338 (main idea information in middle of text)
v342 (main idea information in 1st or 2nd sentence or rest of first short paragraph).

Another way to show the similarities between SAT and GRE main idea item results is to compare all the 22 SAT correlations referred to above with the corresponding 22 GRE correlations. The correlation of these two sets of correlations is significant ($r = .65$, $p < .002$, 2-tailed). Another way to determine the similarity of the two sets of correlations is to compare just the algebraic sign of the 22 SAT correlations with the algebraic sign of the 22 GRE correlations. Seventeen of the 22 correlations are in the same direction which is significant ($p = .016$, 2-tailed, sign test).

This suggests that many of the results reported here appear to be replicable findings, at least for the main idea items.

Conclusion

In this study we have been primarily interested in determining how well main idea reading item difficulty can be accounted for by a set of predictors which reflect the contribution of the text structure, the item structure and the joint effect of both the text and items. We found that a substantial amount of the variance can be accounted for by a relatively small set of predictors; the range of variance accounted for varied from 46% up to 59% depending upon the particular analysis undertaken. The predictability of a larger set of reading items (n=285) was also explored (this varied from 21% to 29% of the variance). To our knowledge this is one of the few studies to examine the predictability of a relatively large sample of multiple-choice reading items (n=285) using a wide selection of predictor variables.

Within this broader concern we have also focused upon a small set of hypotheses so as to more clearly come to terms with a number of claims that have been made in the scholarly literature concerning reading comprehension and the adequacy of reading comprehension tests per se. In particular Goodman (1982) has complained that many of the experimental studies of comprehension have focused on just one or two variables at a time; he questions whether these separate studies taken together necessarily build up our understanding of how full comprehension of text takes place. A related concern has questioned whether the often highly artificial texts studied in the experimental literature will necessarily clarify how more naturalistic texts are comprehended. Finally Royer (1990) and Katz et al. (1990) have questioned whether multiple-choice reading tests can be considered appropriate tests of passage comprehension in light of the fact that item content alone (in the absence of the reading passage) can be demonstrated to lead to correct answers above chance levels of guessing.

In response to these several concerns, we have framed a number of hypotheses meant to put into clearer perspective the viability of multiple-choice reading comprehension tests, here exemplified by the SAT reading passages and their associated items. Since many of the scored variables deal with text content similar to those of concern in the experimental literature and since the SAT reading passages are adaptations of prose from naturalistic sources (book passages, magazines, etc.) we reasoned that the successful prediction of reading item difficulty would allow us to draw several important conclusions. These conclusions were framed as four hypotheses.

The first hypothesis asserts that multiple-choice items will be sensitive to a similar set of variables as have been found to be important in studying comprehension processes in the experimental literature. The evidence

generally was interpreted to support many of the categories detailed under Hypothesis 1 primarily for the text variables. This was interpreted to mean that multiple-choice response formats yield similar results to those found in the more controlled experimental studies. Hence Royer's (1990) claim that multiple-choice tests do not measure passage comprehension has been called into question.

A second hypothesis asserts that many of the significant variables will be found to jointly influence reading item difficulty. Stepwise regression results for main idea items allowed us to conclude that there was considerable evidence that many of the different categories of variables studied do jointly account for reading item difficulty. This result was further interpreted as a response to Goodman's (1982) concern that since many of the experimental studies involve just one or two variables at a time, this may not be sufficient to guarantee that these variables when jointly studied will provide any cumulative new information about reading comprehension difficulty. Our results appear to suggest that in fact many of the different categories of variables do provide independent predictive information; hence the few variables studied across disparate studies do in fact jointly combine so as to increase our understanding of what influences comprehension difficulty. A related set of analyses using a large number (n=244) of GRE reading items (Freedle & Kostin, 1990) further confirmed the viability of this demonstration.

The fact that the SAT passages were selected from naturalistically occurring passages was further interpreted as evidence that the predictive success of many of the text variables found here to predict the difficulty of items associated with these more naturalistic passages are similar to those variables found to predict the difficulty of artificially constructed materials (as is true of many sentences and/or passages in the experimental literature). Hence there do not seem to be any large differences between studies using naturalistic versus artificially constructed materials in terms of their adequacy to study the factors that influence comprehension difficulty. A similar result was obtained with our analyses of GRE data (see Freedle & Kostin, 1991); since these GRE passages are also developed from naturalistically occurring prose passages, this again indicates that the distinction between artificially constructed materials and naturalistic ones is not that great in terms of assessing factors that influence reading comprehension.

A third hypothesis (stated as four variants) dealt with the implications of several studies which support the idea that item variables should account for either all the item difficulty variance or at least should account for

more variance than do text and text associated variables. Hierarchical regression analyses, as expected, did not support this conjecture since most analyses indicated that text plus text associated variables are better predictors than are just item variables. This result casts doubt on some criticisms of multiple-choice reading tests made recently by Royer (1990) and Katz et al. (1990).

An additional hypothesis examined whether the several positions of relevant main idea text information for correctly answering main idea items were related to item difficulty. We found general confirmation of the following nature. Main idea items are facilitated when the relevant key text information occurs early in the text; main idea items become more difficult when relevant information is located in the middle parts of a text. Similarly when all item types are combined (i.e., main ideas plus inferences plus explicit statement items) there is additional evidence that the relative location of even main idea information affects the difficulty of all item types. In general these findings were interpreted as generalizations of earlier empirical work by Kieras (1985) and by Hare et al. (1989). That is, it appears that multiple-choice tests reflect a similar locational effect, and since many of the SAT passages involve multiple paragraphs, it appears that the Kieras (1985) finding generalizes to multiple paragraphs. However it is not immediately clear whether our current data generalize to nontechnical passages (i.e., primarily passages with abstract content). [This issue is explored in greater detail in Appendix C.]

Future directions. Future work should plan to expand the list of text and item predictors along more theoretical lines as suggested by the comparative analyses of Abelson and Black (1986). A more integrative theoretical account of how text processing is assessed by multiple-choice tests is also needed; this should attempt to unite a psychometric model such as that suggested by Embretson and Wetzel (1987) with a text processing approach suggested by Abelson and Black (1986). Once such a higher-level theory is suggested, an attempt can then be made to select only item and text variables which are specifically tied to this theory.

References

Anderson, R. A., & Davison, A. (1988). Conceptual and empirical basis of readability formulas. In A. Davison & G. Green, Linguistic complexity and text comprehension: readability issues reconsidered (pp. 23-53). Hillsdale, N.J.: L. Erlbaum

Abelson, R. P., & Black, J. B. (1986). Introduction. Chapter in J. A. Galambos, R. P. Abelson, & J. B. Black (Eds.), Knowledge structures (pp. 1-18). Hillsdale, N.J.: L. Erlbaum Associates.

Abrahamsen, E., & Shelton, K. (1989). Reading comprehension in adolescence with learning disabilities: semantic and syntactic effects. Journal of Learning Disabilities, 22, 569-572.

Bever, T. G., & Townsend, D. (1979). Perceptual mechanisms and formal properties of main and subordinate clauses. In W. Cooper & E. Walker (Eds.), Sentence processing (pp. 159-226). Hillsdale, N.J.: L. Erlbaum

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. Psychological Review, 82, 45-73.

Clark, H. H., & Haviland, S. (1977). Comprehension and the given-new contract. Chapter in R. Freedle (Ed.), Discourse production and comprehension (pp. 1-40). Norwood, N.J.: Ablex Corp.

Cohen, J., & Cohen, P. (1983). Applied multiple regression--correlational analysis for the behavioral sciences. Hillsdale, N.J.: L. Erlbaum & Assoc.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. Reading Research Quarterly, 16, 486-514.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11, 175-193.

Ervin-Tripp, E. (1964). An analysis of the interaction of language, topic and listener. American Anthropologist, 66, 86-102.

Freedle, R. (1990). Evidence for the construct validity of the SAT reading comprehension items. Paper submitted to Psychological Science.

Freedle, R., & Duran, R. (1979). Sociolinguistic approaches to dialogue with suggested applications to cognitive science. Chapter in R. Freedle (Ed.), New directions in discourse processing (pp. 197-206). Norwood, N.J.: Ablex Corp.

Freedle, R., Fine, J., & Fellbaum, C. (1981). Predictors of good and bad essays. Paper presented at the annual Georgetown University Roundtable on languages and linguistics. Washington, D.C.

Freedle, R., & Kostin, I. (1990). The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: main idea, inference and explicit statement items. Final Report for GRE project. Princeton, N.J.: Educational Testing Service.

Goodman, K. (1982). Language and literacy: The selected writings of Kenneth S. Goodman (Vols 1 & 2). F. Gollasch (Ed.) Boston: Routledge & Kegan Paul.

Graves, M. (1986). Vocabulary learning and instruction. In E. Rothkopf (Ed.), Review of research in education, (pp. 49-90) Vol. 13. Washington, D. C.: American Educational Research Association.

Grimes, J. (1975). The thread of discourse. The Hague: Mouton.

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. Journal of Educational Psychology, 79, 220-227.

Hare, V., Rabinowitz, M., & Schieble, K. (1989). Text effects on main idea comprehension. Reading Research Quarterly, 24, 72-88.

Hites, R. W. (1950). The relation of readability and format to retention in communication. Unpublished doctoral dissertation. Ohio State University.

Hymes, D. (1962). The ethnography of speaking. In T. Gladwin & W. Sturtevant (Eds.), Anthropology and human behavior. Washington, D.C.: Anthropological Society of Washington.

Johnston, P. (1983). Reading comprehension assessment: a cognitive basis. Newark, N.J.: The International Reading Assoc.

Just, M. A., & Carpenter, P. A. (1987). The psychology of reading and language comprehension. Boston: Allyn and Bacon, Inc.

Katz, S., Lautenschlager, G., Blackburn, A., & Harris, F. (1990). Answering reading comprehension items without passages on the SAT. Psychological Science, 1, 122-127.

Kieras, D. E. (1985). Thematic processes in the comprehension of technical prose. Chapter in B. Britton & J. Black (Eds.), Understanding expository text (pp. 89-108). Hillsdale, N.J.: L. Erlbaum Associates.

Kintsch, W. (1974). The representation of meaning in memory. Hillsdale, N.J.: L. Erlbaum.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Klare, G. (1974-1975). Assessing readability. Reading Research Quarterly, 10, 62-102.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. Cognitive Psychology, 9, 111-151.

Meyer, B. (1975). The organization of prose and its effects on memory. Amsterdam: North-Holland.

Meyer, B. (1985). Prose analysis: purposes, procedures and problems. Chapter in B. Britton & J. Black (Eds.), Understanding expository text (pp. 269-304). Hillsdale, N.J.: L. Erlbaum.

Meyer, B., & Freedle, R. (1984). The effects of different discourse types on recall. American Educational Research Journal, 21, 121-143.

Paivio, A. (1971). Imagery and verbal processes. N.Y.: Holt, Rinehart & Winston.

Royer, J. (1990). The sentence verification technique: a new direction in the assessment of reading comprehension. In S. Legg & J. Algina (Eds.), Cognitive assessment of language and math outcomes (pp. 144-191). Norwood, N.J.: Ablex.

Appendix A

Glossary of Variables with a Brief Description

The variables listed below are presented in ascending numerical order for ease of scanning. The variables listed in the materials section of this report have been grouped according to their categorization as text, text by item, or item variables which disrupt the numerical sequencing.

v3 Correct option, position of correct answer among 5 options
v4 Information in first sentence present throughout text, 3= maximum coherence; 0= minimum coherence.
v5 Equated delta
v6 Main idea is in last short paragraph (< 100 words).
v9 R Biserial
v11 Number of paragraphs in passage
v12 Number words in passage
v13 Number of sentences in passage
v14 Number of words in stem
v15 Number of words in correct option
v16 Number of words in all incorrect options
v17 Number of words with three or more syllables in first 100 passage words
v18 Number of words in first paragraph
v19 Number of words in longest paragraph
v31 One of the natural sciences
v32 The second natural science
v33 One of the nontechnical fields
v34 The second nontechnical field
v35 The third nontechnical field--argumentation
v36 Narratives (excluded from this study of expositions)
v37 Main idea information is not explicitly present in text
v39 Main idea information is in rest of first short (< 100 words) paragraph (this does not include coding of 1st or 2nd sentence)
v40 Main idea information in first sentence, paragraph two
v41 Main idea information is in last sentence of passage
v42 Number of sentences in first paragraph
v43 Number of sentences in longest paragraph
v44 Concreteness of passage (1= yes, 0=abstract)
v45 Grimes code: list (or describe) rhetorical organization
v46 Grimes code: causal rhetorical organization

v47 Grimes code: sum of two kinds of comparatives (see v337 which is compare-argumentative and v339 is compare alternative).
v48 Grimes code: problem-solution
v50 Percent of fronted clauses, paragraph beginnings only
v51 Frequency of fronted clauses, paragraph beginnings only
v52 Percent of fronted clauses, total text
v53 Frequency of fronted clauses, total text
v54 Frequency of combinations of fronted material in same clause
v55 Frequency of deferred foci
v56 Frequency of text questions
v57 Number of longest run of fronted consecutive clauses
v58 Passage is "science excerpt" (1= yes, 0 = no)
v59 Passage is "about science" (1= yes, 0 = no)
v60 Main idea item
v61 Inference item
v62 Explicit statement item
v63 Frequency pronoun referentials, within independent clauses
v64 Frequency pronoun referentials, across independent clauses
v65 Frequency pronoun referentials, external text referent
v66 Sum of v63+v64+v65
v67 Text, simple negations (prefixes, suffixes, negative adverbs)
v68 Stem, use of hedges (probably, maybe)
v69 Stem, full question (1= incomplete sentence, 0= full sentence)
v70 Stem, use of negatives (prefixes, suffixes, negative adverbs)
v71 Stem, use of fronting
v72 Stem, sum of pronoun referentials to text, stem or options
v73 Stem, any specific reference to text lines or paragraphs
v75 Correct option, negatives (prefixes, suffixes, negative adverbs)
v76 Correct option, use of fronting (1= yes, 0=no)
v77 Correct option, use of pronoun referentials
v78 Incorrect options, use of negatives
v79 Incorrect options, frontings
v80 Incorrect options, pronoun referentials to text stem or same option
v86 Main idea info. located in first text sentence
v87 Main idea info. located in second text sentence
v88 Percent low ability examinees passing item (v88 through v94 are used only for correlations)
v89 Average number of words per sentence
v90 Average number words per paragraph
v91 Percent 2nd lowest ability examinees passing item
v92 Percent middle ability examinees passing item

v93 Percent 2nd highest ability examinees passing item
v94 Percent highest ability examinees passing item
v96 Average sentence length of first paragraph
v97 Average sentence length of longest paragraph
v100 Text, the natural sciences (v31 and v32)
v337 Grimes rhetorical structure: compare-adversative
v338 Main idea information is in middle of text (not beginning or end)
v339 Grimes: compare-alternative
v342 Main idea information (sum of v86, v87, v39).
v388 z-score of v88 (this variable was used in all regression analyses for low ability examinees)
v394 z-score of v94 (this variable was used in all regression analyses for high ability examinees)

Appendix B

This appendix contains two supplementary tables (Tables 8 & 9) including the means, standard deviations, and correlations for all the variables studied.

Table 8

Means and Standard Deviations
for All Predictor Variables and their Correlations
with Equated Delta and each of Five Ability Levels (percent correct for each level) for 110 Main Idea Items

Var.	Mean	SD	Correlation of Variable with Eq. Delta & each of Five Ability Levels					
			Eq. Delta (uncorrected)	Low	2nd Low	Mid.	2nd Hi	High
v3	2.94	1.36	.19	-.15	-.16	-.18	-.18	-.17
v4	2.71	0.68	-.24	.16	.18	.22	.25	.27
v5	10.79	2.01	1.00	-.89	-.95	-.97	-.96	-.90
v6	0.11	0.31	-.04	.06	.08	.07	.03	-.02
v11	3.20	1.30	-.03	.09	.06	.04	.05	.02
v12	353.04	95.70	15	-.08	-.13	-.14	-.14	-.14
v13	15.16	5.02	07	-.01	-.06	-.09	-.08	-.13
v14	9.38	2.53	18	-.20	-.21	-.17	-.13	-.11
v15	8.65	3.33	15	-.09	-.12	-.16	-.14	-.13
v16	34.14	11.96	09	-.09	-.08	-.09	-.07	-.04
v17	17.15	5.55	-.05	.03	.01	.00	-.02	.03
v18	120.23	61.30	.22	-.23	-.24	-.25	-.23	-.18
v19	161.64	53.66	.27	-.27	-.29	-.26	-.27	-.25
v33	0.27	0.45	-.05	-.02	.02	.09	.11	.10
v34	0.18	0.39	.14	-.07	-.09	-.10	-.14	-.17
v35	0.26	0.44	.39	-.47	-.45	-.43	-.37	-.29
v37	0.26	0.44	.09	-.04	-.05	-.05	-.06	-.06
v39	0.14	0.34	-.20	.21	.22	.19	.19	.13
v40	0.17	0.38	-.17	.06	.13	.16	.18	.21
v41	0.06	0.25	.03	-.16	-.11	-.06	.01	.05
v42	4.91	2.54	.20	-.22	-.22	-.24	-.22	-.21
v43	6.75	2.28	.28	-.28	-.30	-.30	-.29	-.32
v44	0.48	0.50	-.54	.55	.56	.57	.51	.47
v45	0.24	0.42	-.28	.25	.28	.29	.29	.27
v46	0.37	0.47	-.10	.06	.08	.09	.09	.07
v47	0.27	0.42	.39	-.36	-.39	-.40	-.38	-.30
v48	0.12	0.32	.00	.06	.04	.01	-.02	-.09
v50	0.36	0.30	-.05	.06	.06	.09	.08	.04
v51	1.22	1.12	-.04	.05	.06	.09	.08	.04
v52	0.45	0.15	.06	.02	.03	.04	-.04	-.06
v53	7.25	2.80	.08	.01	-.01	-.01	-.07	-.10

Table 8 (Cont.d)

v54	1.21	1.24	-.02	.03	.01	.03	-.01	-.01
v55	0.71	1.03	.20	-.18	-.19	-.18	-.18	-.20
v56	0.38	1.06	.28	-.27	-.26	-.27	-.25	-.29
v57	2.97	1.43	.05	.00	.02	.02	-.06	-.08
v58	0.35	0.48	-.37	.40	.38	.36	.31	.28
v59	0.09	0.29	.20	-.23	-.25	-.25	-.21	-.12
v63	6.46	3.83	.00	.02	.02	.01	.00	-.02
v64	8.15	6.05	.25	-.21	-.22	-.23	-.23	-.32
v65	2.58	4.06	.14	-.07	-.11	-.15	-.14	-.15
v66	17.19	8.90	.24	-.17	-.20	-.23	-.22	-.30
v67	6.67	4.42	.35	-.29	-.31	-.31	-.33	-.37
v68	0.05	0.23	.08	-.09	-.07	-.05	-.03	-.04
v69	0.62	0.49	.02	.01	.01	-.04	-.09	-.10
v71	0.03	0.16	.16	-.14	-.17	-.16	-.18	-.12
v72	0.85	1.00	.00	-.04	-.03	.03	.08	.09
v75	0.09	0.29	.26	-.15	-.19	-.22	-.24	-.26
v77	0.23	0.46	.06	-.04	-.03	-.06	-.05	-.10
v78	0.43	0.76	.08	-.06	-.09	-.09	-.09	-.08
v80	0.78	1.23	-.04	.05	.06	.05	.07	.01
v86	0.25	0.43	-.21	.24	.21	.19	.18	.18
v87	0.17	0.38	-.13	.17	.16	.14	.12	.09
v89	24.34	5.47	.10	-.10	-.11	-.06	-.06	.00
v90	122.88	45.08	.19	-.22	-.22	-.20	-.18	-.16
v96	25.30	6.93	.07	-.04	-.07	-.04	-.05	.01
v97	24.76	6.59	.07	-.04	-.06	-.01	-.04	.00
v100	0.38	0.49	-.32	.33	.32	.30	.27	.25
v337	0.38	0.49	.38	-.42	-.43	-.40	-.36	-.22
v338	0.20	0.40	.22	-.22	-.25	-.24	-.24	-.21
v339	0.13	0.31	.11	-.03	-.06	-.10	-.12	-.12
v342	0.55	0.85	-.25	.29	.27	.23	.22	.18

a

A correlation = .19 is significant at $p < .05$, 2-tailed.
A correlation = .24 is significant at $p < .01$, 2-tailed.
A correlation = .31 is significant at $p < .001$, 2-tailed.
The algebraic sign for the equated delta correlations have not been reversed in this table. A positive correlation for the ability groups (calculated using the percent pass scores only) facilitates performance. A negative correlation for equated delta however facilitates performance. The reader should note that the correlations for high and low ability groups presented in

this table differ slightly from those presented in table 1 of this report (which used a z-score transformation prior to computing the correlation).

b

The following variables were omitted from further analysis due to their low frequency of occurrence among the main idea items (i.e., had fewer than 3 entries):

v70 (stem, use of simple negations)
v73 (stem, references made to text lines or paragraphs)
v76 (correct, use of fronting)
v79 (incorrect, use of fronting)

Table 9

Means and Standard Deviations of each Predictor Variable for All Reading Item Types (n=285) and Correlations of each Variable with Six Criterion Variables: Equated Delta and each of Five Ability Levels

[a]

			Equated Delta and each of Five Ability Levels					
Var.	Mean	SD	Eq.Delta	Low	2ndLo	Mid	2ndHi	High
v3	2.95	1.40	.05	-.04	-.03	-.04	-.03	-.05
v4	2.68	0.73	-.11	.07	.07	.09	.11	.13
v5	11.29	2.20	1.00	-.91	-.95	-.97	-.96	-.90
v6	0.12	0.33	.03	.03	.01	-.01	-.03	-.06
v11	3.32	1.26	-.03	.03	.02	.05	.05	.06
v12	364.97	91.08	.08	-.09	-.10	-.09	-.07	-.04
v13	15.82	5.00	-.01	.02	.00	-.01	.01	.01
v14	13.82	5.28	.21	-.22	-.22	-.22	-.19	-.20
v15	8.44	6.55	.10	-.08	-.09	-.10	-.11	-.14
v16	31.22	13.13	.07	-.08	-.08	-.07	-.08	-.08
v17	17.07	5.66	.05	-.09	-.09	-.09	-.07	-.05
v18	118.83	60.52	.13	-.14	-.13	-.16	-.14	-.11
v19	162.44	51.48	.13	-.15	-.16	-.16	-.15	-.13
v33	0.29	0.46	.03	-.06	-.04	.00	.01	.00
v34	0.18	0.38	.11	-.06	-.07	-.10	-.13	-.12
v35	0.27	0.44	.15	-.21	-.21	-.20	-.15	-.11
v37	0.28	0.45	.06	-.07	-.04	-.03	.00	.01
v39	0.14	0.34	-.20	.19	.20	.21	.19	.17
v40	0.15	0.36	-.14	.10	.10	.13	.14	.15
v41	0.06	0.24	.06	-.10	-.10	-.09	-.06	-.04
v42	4.84	2.42	.06	-.06	-.05	-.09	-.07	-.07
v43	6.80	2.19	.08	-.09	-.10	-.12	-.11	-.09
v44	0.46	0.50	-.35	.34	.36	.37	.34	.31
v45	0.25	0.42	-.10	.08	.10	.11	.12	.08
v46	0.35	0.46	-.09	.06	.08	.08	.07	.08
v47	0.28	0.43	.24	-.20	-.24	-.25	-.22	-.15
v48	0.13	0.32	-.06	.08	.07	.06	.03	-.02
v50	0.36	0.30	-.04	.03	.05	.06	.06	.03
v51	1.27	1.12	-.06	.04	.07	.09	.09	.06
v52	0.45	0.15	-.04	.05	.09	.08	.07	.06
v53	7.43	2.80	-.04	.04	.06	.06	.05	.06
v54	1.23	1.24	-.01	-.01	.00	.00	-.01	.00

Table 9 (Cont'd.)

v55	0.71	1.02	.05	-.06	-.04	-.04	.00	-.03
v56	0.39	1.03	.08	-.08	-.07	-.07	-.04	-.04
v57	2.93	1.44	-.05	.03	.08	.09	.08	.08
v58	0.34	0.47	-.18	.20	.20	.18	.15	.11
v59	0.09	0.29	.10	-.10	-.12	-.13	-.11	-.06
v60	0.37	0.49	-.18	.19	.19	.19	.16	.15
v61	0.34	0.47	.25	-.23	-.24	-.25	-.25	-.27
v62	0.27	0.47	-.07	.03	.04	.07	.09	.12
v63	6.51	3.69	-.01	-.03	-.01	-.01	.00	.01
v64	8.36	6.16	.07	-.04	-.06	-.06	-.05	-.08
v65	2.58	4.13	.00	.03	.01	-.02	-.01	-.01
v66	17.45	8.89	.04	-.03	-.04	-.06	-.04	-.06
v67	6.95	4.45	.15	-.14	-.14	-.14	-.13	-.13
v68	0.06	0.24	.13	-.11	-.11	-.11	-.13	-.15
v69	0.68	0.47	.03	-.01	-.02	-.04	-.05	-.05
v71	0.39	0.52	.06	-.07	-.08	-.06	-.05	-.04
v72	0.86	0.96	.00	-.05	-.02	.01	.04	.06
v75	0.15	0.39	.16	-.15	-.16	-.17	-.17	-.14
v77	0.30	0.56	-.01	.01	.02	.01	.01	-.03
v78	0.68	1.05	.21	-.19	-.22	-.23	-.23	-.23
v80	1.04	1.51	.01	-.02	.00	-.02	.01	-.03
v86	0.23	0.42	-.06	.09	.07	.03	.01	.00
v87	0.15	0.36	-.08	.09	.11	.08	.07	.05
v89	24.10	5.24	.10	-.11	-.10	-.07	-.08	-.05
v90	121.55	43.25	.12	-.14	-.13	-.14	-.13	-.11
v96	25.24	6.84	.13	-.13	-.14	-.11	-.11	-.06
v97	24.71	6.46	.08	-.09	-.08	-.05	-.06	-.05
v100	0.37	0.48	-.17	.16	.16	.16	.14	.11
v337	0.13	0.34	.20	-.22	-.25	-.23	-.19	-.10
v338	0.20	0.40	.03	-.01	-.03	-.04	-.05	-.06
v339	0.15	0.33	.10	-.03	-.06	-.09	-.09	-.09
v342	0.52	0.82	-.15	.17	.17	.14	.12	.10

a

A correlation = .12 is significant at $p < .05$, 2-tailed

A correlation = .15 is significant at $p < .01$, 2-tailed

A correlation = .20 is significant at $p < .001$, 2-tailed.

A positive correlation for the ability groups indicates a facilitating effect on percent correct (all correlations for ability groups used percent pass only. The algebraic sign of equated delta correlations have not been reversed; here a negative correlation indicates a facilitating effect.

The reader should note that the correlations reported here differ slightly from those entered in Table 4 of this report (which used a z-score transformation prior to computing the correlation.

Appendix C

Mean percent correct scores for the Kieras-type scores reflecting positions of the main idea in abstract versus concrete type passages for high and low ability examinees

There are eight positions which were targeted for the location of the main idea in our 110 passages. These varied from early in the text, the opening two sentences, to the last sentence in the text. We present these detailed results below for concrete (primarily technical) and abstract (primarily nontechnical) passages in order to provide another way to evaluate whether it is legitimate to generalize Kieras' (1985) findings. Kieras (1985) studied only technical prose. The results below will allow us a more careful examination of whether his findings generalize to the abstract passage which are primarily nontechnical in content.

To keep the results clear-cut, only those passages that had a unique Kieras code were selected for analysis (some passages had several positions which were judged to be a statement, or partial statement, of the main idea item--as reflected by the keyed correct answer to a main idea item associated with that passage). These passages with multiple entries of the Kieras type were excluded from the analysis below. This resulted in 80 passages being selected for analysis having just one main idea item per passage. The main body of this reports analyzes v342 which includes the sum of codes v86, v87 and v39; this was combined because it increased the size of the correlation of the Kieras codes with item difficulty.

Effect of concrete passages (mostly technical) on main idea correctness

Table I presents the results of the effect on main idea difficulty when the main idea is located in different text positions.

Table I

Mean Percent Correct Main Ideas by Text Position
for Concrete (mostly Technical) Passages
(n=35 passages, 35 main idea items)

Position		Mean Percent Correct of all Main Idea Items at a given text position for two ability levels: Ability Level of Examinees: High	Low
v86,	1st sentence, 1st paragraph	92	51
v87,	2nd sentence, 1st paragraph	94	58
v39,	First short (< 100 words) paragraph (not including v86,v87)	89	50
v40,	1st sentence, 2nd paragraph	91	42
v338,	Middle of text	80	28
v6,	In last short (< 100 words.) paragraph	87	52
v41,	Last sentence of text	85	19
v37,	Not clearly stated in text	85	40

Results for concrete passages (mostly technical):

From a cursory examination of Table I we see that the opening sentences have the highest mean percentage main idea items correct for the high ability people; this is also true for v87 of the low ability examinees, with v86 being the third highest percent correct. These results are interpreted as consistent with what one would expect from the Kieras (1985) study. That is, since most students think the main idea is often in the opening sentences, when it in fact is located there (as determined by the keyed answer in our multiple-choice data) the examinees often get such main idea items correct. We also see that when main idea information is located in the middle of the passages they tend to yield among the lowest mean percent correct for both high and low ability groups. This is also consistent with our interpretation of the Kieras (1985) study. The high and low ability examinees appear to differ substantially on how well they recognize main idea information that appears only in the last sentence of the text. The high group does quite well, whereas the low ability group falls below chance (20%) level. The only real surprise in

this table is that when no explicit information is present in the surface text (v37) concerning the main idea, the high and low ability people still do relatively well.

Effect of abstract passages (primarily nontechnical) on main idea correctness

Table II presents the results for the effect of different locations for the main idea of a passage as a function of the abstract passages.

Table II
Mean Percent Correct Main Ideas by Text Position
for Abstract Passages
(45 passages, 45 main idea items)

Position		Mean Percent Correct of all Main Idea Items at a given text position for two ability levels: Ability Level of Examinees High	Low
v86,	1st sentence, 1st paragraph	71	28
v87,	2nd sentence, 1st paragraph	--	--
v39,	First short (< 100 words) paragraph (not including v86,v87)	78	23
v40,	First sentence, 2nd paragraph	77	24
v338,	Middle of text	71	22
v6,	In last short (< 100 words) paragraph	70	24
v41,	Last sentence of text	78	18
v37,	Not clearly stated in text	72	24

-- means no main idea items fell into this particular text position for the abstract texts.

Results for abstract passages (most nontechnical)

A quick look at the mean percent corrects for the abstract (generally nontechnical) passages presented in Table II indicates an unexpected result. The high ability people do not seem to be very sensitive to the relative location of the main idea information in the text. Neither do the low ability people. In fact most of the entries for the low ability examinees appears to

be very close to chance (20% correct) performance. This result suggests that our interpretation of the Kieras-type findings for technical, single paragraph prose may not readily generalize to nontechnical, multiparagraph prose--at least not for the low ability examinees.

Comparing across tables I and II, we see several things that they have in common. The high ability examinees consistency perform better than low ability examinees for each Kieras-type position of the main idea, $p < .01$ in each case, using a non-parametric sign test.

Of greater interest, we see that the mean percent passing at a given text position for the concrete passages is significantly higher than for the same position for abstract passages ($p = .016$, 1-tail sign test) for high ability examinees. The same significance level ($p = .016$, 1-tail sign test) applies for the low ability examinees across compared across abstract and concrete texts. Thus concrete main idea items are significantly easier than abstract main idea items when surface text position is controlled for.

Finally, we see that only the main ideas for concrete passages appear to support our generalization of the Kieras results. This is not surprising since our concrete passages generally represent technical types of prose; the fact that the multi-paragraph contexts of our data do not differ that much from the Kieras type results supports the idea that the length of the passages does not interfere with our generalization.

However, the apparent failure to generalize to the abstract passages represents a new finding in two senses: it suggests that the ability groups differ substantially for abstract passages and it suggests that the examinees are processing these texts differently for main idea information inasmuch as relative text location does not affect percentage correct in the same way that it does for the concrete passages. This clearly requires further studies to clarify the nature of these processing differences.